

# Chapitre Premier

## La Géométrie des Moindres Carrés

### 1.1 INTRODUCTION

La méthode d'estimation la plus communément utilisée en économétrie, et à bien des égards la plus importante, est celle des **moindres carrés**. Il est utile de distinguer deux variétés de moindres carrés, les **moindres carrés ordinaires**, ou **OLS**, et les **moindres carrés non linéaires**, ou **NLS**. Dans le cas des OLS, l'équation de la régression qui doit être estimée est linéaire en tous les paramètres, alors que dans le cas des NLS, elle est non linéaire en au moins un des paramètres. Les estimations par OLS peuvent être obtenues par un calcul direct de plusieurs manières (consulter la Section 1.5), alors que les estimations par NLS réclament l'utilisation de procédures itératives (consulter le Chapitre 6). Dans ce chapitre, nous traiterons exclusivement des OLS, puisque la compréhension de la régression linéaire est essentielle à la compréhension de tout ce qui suit dans le livre.

Il faut faire une distinction importante entre les propriétés numériques et les propriétés statistiques des estimations obtenues par l'utilisation des OLS. Les **propriétés numériques** sont celles qui apparaissent comme une conséquence de l'utilisation des OLS, sans se soucier de la façon dont les données ont été générées. Comme ces propriétés sont numériques, elles peuvent toujours être vérifiées par un calcul direct. Un exemple bien connu est que la somme des résidus des OLS est nulle lorsque les régresseurs contiennent un terme constant. D'autre part, les **propriétés statistiques** sont celles qui ne subsistent que sous certaines hypothèses sur la façon dont les données ont été générées. Celles-ci ne peuvent jamais être vérifiées avec exactitude, bien que dans certains cas elles puissent être testées. Un exemple est la proposition bien connue que les estimations par OLS sont non biaisées, dans certaines circonstances.

La distinction entre des propriétés qui sont exactes numériquement et des propriétés statistiques est à l'évidence fondamentale. Dans le but d'illustrer cette distinction le plus clairement possible, nous ne discuterons dans ce présent chapitre que des premières. Nous n'étudierons les OLS que comme un système d'estimation, sans introduire formellement un quelconque modèle statistique (bien que nous aurons l'occasion de discuter de plusieurs modèles qui sont surtout intéressants dans le contexte des modèles de régression



linéaire). Aucun modèle statistique ne sera introduit avant le Chapitre 2, où nous entamerons une discussion sur les **modèles de régression non linéaire**, dont les **modèles de régression linéaire** ne sont bien sûr que des cas particuliers.

En annonçant que nous étudierions les OLS en tant que système d'estimation, nous ne prétendons pas discuter des algorithmes informatiques pour le calcul des estimations par OLS (toutefois, nous en discuterons dans la Section 1.5). Au contraire, nous voulions dire par là que nous discuterions des propriétés numériques des moindres carrés ordinaires, et en particulier de l'interprétation géométrique de celles-ci. Toutes les propriétés numériques des OLS peuvent être interprétées à l'aide de la géométrie Euclidienne. Cette interprétation géométrique se révèle souvent être remarquablement élémentaire, impliquant à peine plus que le Théorème de Pythagore et la trigonométrie enseignée au lycée, dans un contexte d'espaces vectoriels à dimension finie. Pourtant, l'aperçu qu'offre cette approche est très appréciable. Une fois que l'on a acquis une connaissance profonde de la géométrie en cause dans les moindres carrés ordinaires, il est possible de s'épargner plusieurs lignes fastidieuses d'algèbre par l'utilisation d'un argument géométrique. De plus, comme nous espérons que le reste du livre l'illustrera, comprendre les propriétés géométriques des OLS est fondamental autant pour comprendre les modèles non linéaires de tous types, que pour comprendre les modèles de régression linéaire.

## 1.2 LA GÉOMÉTRIE DES MOINDRES CARRÉS

Les éléments essentiels d'une régression linéaire sont la **régressande**  $\mathbf{y}$  et une matrice de **régresseurs**  $\mathbf{X} \equiv [\mathbf{x}_1 \dots \mathbf{x}_k]$ . La régressande  $\mathbf{y}$  est un vecteur de dimension  $n$ , et la matrice des régresseurs  $\mathbf{X}$  est une matrice de dimension  $n \times k$ , dont chaque colonne  $\mathbf{x}_i$  est un vecteur de dimension  $n$ . La régressande  $\mathbf{y}$  et chaque régresseur, de  $\mathbf{x}_1$  à  $\mathbf{x}_k$ , peuvent être interprétés comme des points dans un **espace Euclidien de dimension  $n$** ,  $E^n$ . Les  $k$  régresseurs, pourvu qu'ils soient linéairement indépendants, **engendrent** un **sous-espace** de dimension  $k$  de  $E^n$ . Nous désignerons ce sous-espace par  $\mathcal{S}(\mathbf{X})$ .<sup>1</sup>

Le sous-espace  $\mathcal{S}(\mathbf{X})$  est constitué par tous les points  $\mathbf{z}$  appartenant à  $E^n$  tels que  $\mathbf{z} = \mathbf{X}\boldsymbol{\gamma}$  pour  $\boldsymbol{\gamma}$  donné, où  $\boldsymbol{\gamma}$  est un vecteur de dimension  $k$ . A proprement parler, nous devrions nous référer à  $\mathcal{S}(\mathbf{X})$  en tant que sous-espace engendré par les colonnes de  $\mathbf{X}$ , mais de manière moins formelle, nous nous y référerons en tant que sous-espace engendré par  $\mathbf{X}$ . La dimension de  $\mathcal{S}(\mathbf{X})$  est toujours égale à  $\rho(\mathbf{X})$ , le **rang** de  $\mathbf{X}$  (c'est-à-dire le nombre de colonnes de  $\mathbf{X}$

<sup>1</sup> La notation  $\mathcal{S}(\mathbf{X})$  n'est pas une notation classique, et il n'y a pas de notation classique avec laquelle nous nous sentons à l'aise. Nous croyons que cette notation a de nombreuses qualités, aussi la recommandons nous et l'utiliserons dès à présent.

qui sont linéairement indépendantes. Nous supposons que  $k$  est strictement inférieur à  $n$ , ce qu'il est raisonnable de faire dans presque tous les cas concrets. Si  $n$  était plus petit que  $k$ , alors  $\mathbf{X}$  ne pourrait pas être de plein rang  $k$ .

Un espace Euclidien n'est pas défini tant que l'on a pas spécifié un **produit intérieur**. Dans le cas précis qui nous intéresse, le produit intérieur est ainsi appelé **produit intérieur naturel**. Le produit intérieur naturel de deux points quelconques de  $E^n$ , disons  $\mathbf{z}_i$  et  $\mathbf{z}_j$ , peut être noté  $\langle \mathbf{z}_i, \mathbf{z}_j \rangle$ , et il est défini par

$$\langle \mathbf{z}_i, \mathbf{z}_j \rangle \equiv \sum_{t=1}^n z_{it} z_{jt} \equiv \mathbf{z}_i^\top \mathbf{z}_j \equiv \mathbf{z}_j^\top \mathbf{z}_i.$$

Remarquons que le produit intérieur naturel n'est pas le seul qui soit à notre disposition; on pourrait par exemple choisir d'attribuer une pondération positive différente à chaque élément de la somme, comme dans

$$\sum_{t=1}^n w_t z_{it} z_{jt}.$$

Ainsi que nous le verrons dans le Chapitre 9, exécuter une régression linéaire par l'intermédiaire de ce produit intérieur reviendrait à utiliser une forme particulière des moindres carrés généralisés. Pour la suite du livre, et à moins que l'on spécifie le contraire, c'est au produit intérieur Euclidien auquel nous ferons référence lorsque nous parlerons de produit intérieur.

Si un point  $\mathbf{z}$  (qui est bien entendu un vecteur à  $n$  composantes) appartient à  $\mathcal{S}(\mathbf{X})$ , il est toujours possible d'écrire  $\mathbf{z}$  comme une combinaison linéaire des colonnes de  $\mathbf{X}$ , de telle sorte que:

$$\mathbf{z} = \sum_{i=1}^k \gamma_i \mathbf{x}_i = \mathbf{X}\boldsymbol{\gamma},$$

où  $\gamma_1, \gamma_2, \dots, \gamma_k$  sont des scalaires, et  $\boldsymbol{\gamma}$  est un vecteur à  $k$  composantes dont l'élément type est  $\gamma_i$ . Ainsi un vecteur de  $k$  coefficients comme  $\boldsymbol{\gamma}$  identifie n'importe quel point dans  $\mathcal{S}(\mathbf{X})$ . Pourvu que les colonnes de  $\mathbf{X}$  soient **linéairement indépendantes**, il le fait de manière unique. Les vecteurs  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$  sont linéairement indépendants s'il est impossible d'écrire l'un d'entre eux comme une combinaison linéaire des autres.

Si les  $k$  régresseurs ne sont pas linéairement indépendants, alors ils engendreront un sous-espace de dimension  $k'$  (avec  $k'$  plus petit que  $k$ ), où  $k'$  est le nombre maximum de colonnes de  $\mathbf{X}$  qui sont linéairement indépendantes entre elles, c'est-à-dire  $\rho(\mathbf{X})$ . Dans ce cas de figure,  $\mathcal{S}(\mathbf{X})$  sera identique à  $\mathcal{S}(\mathbf{X}')$ , où  $\mathbf{X}'$  est une matrice de dimension  $n \times k'$  constituée des  $k'$  colonnes de  $\mathbf{X}$  linéairement indépendantes. Par exemple, considérons la matrice  $\mathbf{X}$

suivante, qui est de dimension  $6 \times 3$ :

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

Les colonnes de cette matrice ne sont pas linéairement indépendantes puisque

$$\mathbf{x}_1 = .5\mathbf{x}_2 + \mathbf{x}_3.$$

Toutefois, n'importe quelle paire de colonnes est constituée de colonnes linéairement indépendantes, et alors

$$\mathcal{S}(\mathbf{X}) = \mathcal{S}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{S}(\mathbf{x}_1, \mathbf{x}_3) = \mathcal{S}(\mathbf{x}_2, \mathbf{x}_3).$$

Nous venons d'introduire une nouvelle notation:  $\mathcal{S}(\mathbf{x}_1, \mathbf{x}_2)$  désigne le sous-espace engendré par les deux vecteurs  $\mathbf{x}_1$  et  $\mathbf{x}_2$  réunis. Plus généralement, la notation  $\mathcal{S}(\mathbf{Z}, \mathbf{W})$  désignera le sous-espace engendré par les colonnes des matrices  $\mathbf{Z}$  et  $\mathbf{W}$  prises simultanément; ainsi  $\mathcal{S}(\mathbf{Z}, \mathbf{W})$  signifie la même chose que  $\mathcal{S}([\mathbf{Z} \ \mathbf{W}])$ . Notons que, dans bien des cas,  $\mathcal{S}(\mathbf{Z}, \mathbf{W})$  sera un espace de dimension inférieure à la somme des rangs de  $\mathbf{Z}$  et de  $\mathbf{W}$ , si quelques colonnes de  $\mathbf{Z}$  se trouvent déjà dans  $\mathcal{S}(\mathbf{W})$ , et vice versa. Pour le reste du chapitre, et à moins que le contraire ne soit explicitement supposé, nous admettrons pourtant que les colonnes de  $\mathbf{X}$  sont linéairement indépendantes.

La première chose qu'il faut noter à propos de  $\mathcal{S}(\mathbf{X})$  est que l'on peut soumettre  $\mathbf{X}$  à n'importe quelle transformation linéaire qui conserve le rang de  $\mathbf{X}$ , sans modifier en quoi que ce soit le sous-espace engendré par la matrice  $\mathbf{X}$  transformée. Si  $\mathbf{z} = \mathbf{X}\boldsymbol{\gamma}$  et

$$\mathbf{X}^* = \mathbf{X}\mathbf{A},$$

où  $\mathbf{A}$  est une matrice non singulière de dimension  $k \times k$ , alors

$$\mathbf{z} = \mathbf{X}^*\mathbf{A}^{-1}\boldsymbol{\gamma} \equiv \mathbf{X}^*\boldsymbol{\gamma}^*.$$

Ainsi, n'importe quel point  $\mathbf{z}$  qui peut s'exprimer comme une combinaison linéaire des colonnes de  $\mathbf{X}$ , peut aussi bien être écrit comme une combinaison linéaire d'une transformation linéaire de ces colonnes. Nous en concluons que si  $\mathcal{S}(\mathbf{X})$  est un espace engendré par les colonnes de  $\mathbf{X}$ , il doit aussi l'être par les colonnes de  $\mathbf{X}^* = \mathbf{X}\mathbf{A}$ . Cela signifie que nous pourrions donner à cet espace une infinité de noms, dans ce cas  $\mathcal{S}(\mathbf{X})$ ,  $\mathcal{S}(\mathbf{X}^*)$ , ou d'autres encore. Quelques auteurs (dont Seber, 1980; Fisher, 1981) ont à cette occasion adopté une notation avec laquelle le sous-espace que nous avons appelé  $\mathcal{S}(\mathbf{X})$  est

désigné sans référence explicite à  $\mathbf{X}$ . Nous avons évité cette notation “sans coordonnées” parce que cela conduit à obscurcir la relation entre les résultats et la (les) régression (régressions) dont ils sont issus; et parce que dans la plupart des cas, il y a un choix évident de la matrice dont l’espace engendré nous intéresse. Cependant, ainsi que nous aurons l’opportunité de le voir, de nombreux résultats essentiels de la régression linéaire sont “sans coordonnées”, dans le sens où ils ne dépendent de  $\mathbf{X}$  que par l’intermédiaire de  $\mathcal{S}(\mathbf{X})$ .

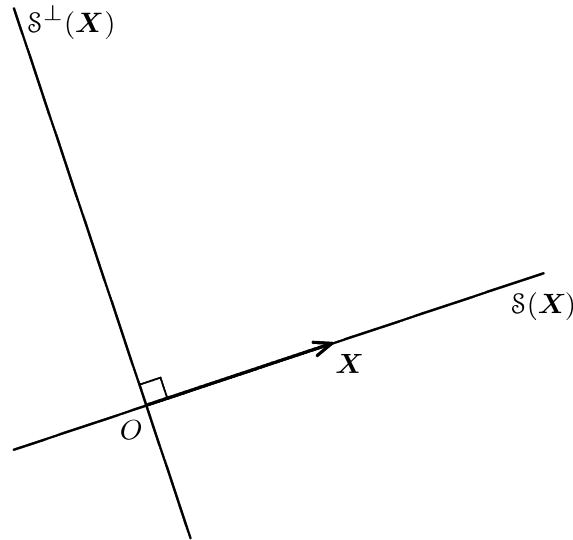
Le **complément orthogonal** de  $\mathcal{S}(\mathbf{X})$  dans  $E^n$ , que l’on note  $\mathcal{S}^\perp(\mathbf{X})$ , est l’ensemble de tous les points  $\mathbf{w}$  appartenant à  $E^n$  pour lesquels on a  $\mathbf{w}^\top \mathbf{z} = 0$ , quel que soit  $\mathbf{z}$  appartenant à  $\mathcal{S}(\mathbf{X})$ . Ainsi, chaque point de  $\mathcal{S}^\perp(\mathbf{X})$  est **orthogonal** à tout point de  $\mathcal{S}(\mathbf{X})$  (deux points sont déclarés orthogonaux si leur produit intérieur est nul). Puisque la dimension de  $\mathcal{S}(\mathbf{X})$  est  $k$ , la dimension de  $\mathcal{S}^\perp(\mathbf{X})$  est  $n - k$ . Il est parfois commode de ne pas se référer à la dimension d’un sous-espace linéaire mais à sa **codimension**. On dira qu’un sous-espace linéaire de  $E^n$  est de codimension  $j$ , si la dimension de son complément orthogonal est  $j$ . Ainsi, dans ce cas,  $\mathcal{S}(\mathbf{X})$  est de dimension  $k$  et de codimension  $n - k$ , et  $\mathcal{S}^\perp(\mathbf{X})$  est de dimension  $n - k$  et de codimension  $k$ .

Avant de discuter de la Figure 1.1, qui illustre ces idées générales, il nous faut faire quelques précisions concernant les conventions géométriques. Le moyen le plus simple de représenter un vecteur à  $n$  composantes, par exemple  $\mathbf{z}$ , dans un diagramme, est de le représenter tout simplement comme un point dans un espace à  $n$  dimensions; mais  $n$  est évidemment limité à 2 ou 3. Il est quelquefois plus intuitif de décrire explicitement  $\mathbf{z}$  comme un vecteur, dans le sens géométrique du terme. Cela se fait en joignant le point  $\mathbf{z}$  à l’origine et en déposant une pointe sur le point. La flèche qui en résulte montre alors graphiquement deux choses à propos du vecteur qui nous intéresse, c’est-à-dire sa **longueur** et sa **direction**. La longueur Euclidienne d’un vecteur  $\mathbf{z}$  est

$$\|\mathbf{z}\| \equiv \left( \sum_{t=1}^n z_t^2 \right)^{1/2} = |(\mathbf{z}^\top \mathbf{z})^{1/2}|,$$

où la notation attire l’attention sur le fait que  $\|\mathbf{z}\|$  est la racine carrée positive de la somme des éléments au carré de  $\mathbf{z}$ . La direction est le vecteur lui-même normalisé, afin d’avoir une longueur unitaire, c’est-à-dire  $\mathbf{z}/\|\mathbf{z}\|$ . L’un des avantages de cette convention d’écriture est qu’en bougeant l’une des flèches, en faisant attention à ne modifier ni sa longueur ni sa direction, la nouvelle flèche obtenue représente le même vecteur, bien que la pointe se situe sur un point différent. Il sera souvent très commode de recourir à ce procédé, aussi adopterons nous cette convention dans la plupart de nos diagrammes.

La Figure 1.1 illustre les concepts dont nous venons de discuter plus tôt pour le cas  $n = 2$  et  $k = 1$ . La matrice de régresseurs  $\mathbf{X}$  possède seulement une colonne de telle sorte qu’elle se représente par un vecteur sur la figure. Par conséquent,  $\mathcal{S}(\mathbf{X})$  est unidimensionnel, et puisque  $n = 2$ ,  $\mathcal{S}^\perp(\mathbf{X})$  est également unidimensionnel. Notons que  $\mathcal{S}(\mathbf{X})$  et  $\mathcal{S}^\perp(\mathbf{X})$  seraient identiques si  $\mathbf{X}$  était



**Figure 1.1** Les espaces  $\mathcal{S}(\mathbf{X})$  et  $\mathcal{S}^\perp(\mathbf{X})$

n'importe quel point sur la ligne droite qu'est  $\mathcal{S}(\mathbf{X})$ , excepté l'origine. Cela illustre le fait que  $\mathcal{S}(\mathbf{X})$  est invariant à toute transformation non singulière de  $\mathbf{X}$ .

Comme nous l'avons vu, n'importe quel point de  $\mathcal{S}(\mathbf{X})$  peut être décrit avec un vecteur de la forme  $\mathbf{X}\boldsymbol{\beta}$  pour un vecteur  $\boldsymbol{\beta}$  donné à  $k$  composantes. Si l'on recherchait le point de  $\mathcal{S}(\mathbf{X})$  qui soit le plus proche d'un vecteur  $\mathbf{y}$  donné, alors le problème qu'il s'agirait de résoudre serait celui d'une minimisation, par rapport à  $\boldsymbol{\beta}$ , de la distance entre  $\mathbf{y}$  et  $\mathbf{X}\boldsymbol{\beta}$ . Minimiser cette distance est bien sûr équivalent à minimiser le carré de cette distance, de telle manière que résoudre

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad (1.01)$$

nous permettra de trouver le point de  $\mathcal{S}(\mathbf{X})$  le plus proche de  $\mathbf{y}$ . La valeur de  $\boldsymbol{\beta}$  solution de (1.01), qui est l'estimation OLS, sera notée  $\hat{\boldsymbol{\beta}}$ .

La distance au carré entre  $\mathbf{y}$  et  $\mathbf{X}\boldsymbol{\beta}$  peut aussi s'écrire comme

$$\sum_{t=1}^n (y_t - \mathbf{X}_t\boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.02)$$

où  $y_t$  et  $\mathbf{X}_t$  sont respectivement le  $t^{\text{ième}}$  élément du vecteur  $\mathbf{y}$  et la  $t^{\text{ième}}$  ligne de la matrice  $\mathbf{X}$ .<sup>2</sup> Puisque l'on fait souvent référence à la différence entre  $y_t$  et  $\mathbf{X}_t\boldsymbol{\beta}$  comme à un **résidu**, cette quantité est généralement appelée la **somme**

<sup>2</sup> Nous nous référons à la  $t^{\text{ième}}$  ligne de  $\mathbf{X}$  en tant que  $\mathbf{X}_t$  plutôt que  $\mathbf{x}_t$  pour éviter la confusion entre les colonnes de  $\mathbf{X}$ , que nous avons déjà notées  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , et ainsi de suite.

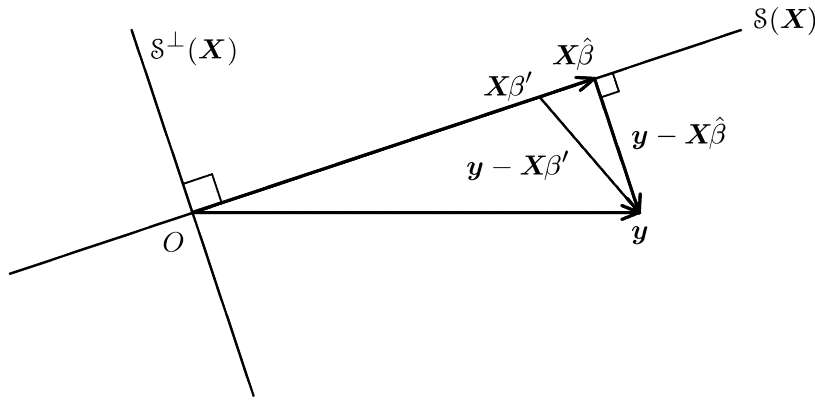


Figure 1.2 La projection de  $\mathbf{y}$  sur  $\mathcal{S}(\mathbf{X})$

des résidus au carré, ou **SSR**. Elle est souvent appelée **somme résiduelle des carrés**, un terme qui se rapproche davantage de la terminologie utilisée pour sa contrepartie, la **somme expliquée des carrés**. Les sigles seraient alors RSS et ESS. Il est dommageable que certains auteurs utilisent le premier sigle pour désigner la somme des carrés de la régression, et le second pour désigner la somme des carrés des erreurs. Alors si l'on utilise les sigles RSS et ESS, il n'est pas du tout clair de repérer le concept que l'on désigne. Lorsque nous ferons appel à SSR et ESS, il ne devra pas y avoir d'ambiguïté quant aux concepts qu'ils désignent.

La géométrie des moindres carrés ordinaires est illustrée dans la Figure 1.2, qui n'est que la Figure 1.1 avec quelques éléments supplémentaires. La régressande est désormais représentée par le vecteur  $\mathbf{y}$ . Le vecteur  $\mathbf{X}\hat{\beta}$ , auquel on fait souvent référence en tant que vecteur de **valeurs ajustées**, est le point de  $\mathcal{S}(\mathbf{X})$  le plus proche de  $\mathbf{y}$ ; notons que  $\hat{\beta}$  est un scalaire dans ce cas précis. Il est évident que la ligne joignant  $\mathbf{y}$  à  $\mathbf{X}\hat{\beta}$  doit former un angle droit avec  $\mathcal{S}(\mathbf{X})$  au point  $\mathbf{X}\hat{\beta}$ . Cette ligne est tout simplement le vecteur  $\mathbf{y} - \mathbf{X}\hat{\beta}$ , translaté de telle manière que son origine est le point  $\mathbf{X}\hat{\beta}$ , et non plus zéro. L'angle droit formé par  $\mathbf{y} - \mathbf{X}\hat{\beta}$  et  $\mathcal{S}(\mathbf{X})$  est la caractéristique clef des moindres carrés. En n'importe quel autre point de  $\mathcal{S}(\mathbf{X})$ , tel que  $\mathbf{X}\beta'$  sur la figure,  $\mathbf{y} - \mathbf{X}\beta'$  ne forme pas d'angle droit avec  $\mathcal{S}(\mathbf{X})$  et en conséquence,  $\|\mathbf{y} - \mathbf{X}\beta'\|$  doit nécessairement être supérieur à  $\|\mathbf{y} - \mathbf{X}\hat{\beta}\|$ .

Le vecteur des dérivées partielles de SSR (1.02) par rapport aux composantes de  $\beta$  est

$$-2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta),$$

qui doit être égal à zéro lorsque l'on est en un minimum. Puisque nous avons supposé que les colonnes de  $\mathbf{X}$  sont linéairement indépendantes, la matrice  $\mathbf{X}^\top\mathbf{X}$  doit être de plein rang. Lorsque l'on combine cet argument avec celui consistant à dire que n'importe quelle matrice de la forme  $\mathbf{X}^\top\mathbf{X}$  est nécessairement définie non négative, il vient que la somme des résidus au carré est une fonction strictement convexe en  $\beta$ , de sorte qu'elle possède un

minimum unique. Ainsi  $\hat{\beta}$  est déterminé de façon unique par les **équations normales**

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}. \quad (1.03)$$

Ces équations normales indiquent que le vecteur  $\mathbf{y} - \mathbf{X}\hat{\beta}$  doit être orthogonal à toutes les colonnes de  $\mathbf{X}$ , et en conséquence, à n'importe quel vecteur se trouvant dans l'espace engendré par ces colonnes. Les équations normales (1.03) sont ainsi une façon de constater algébriquement ce que nous avons vu géométriquement à partir de la Figure 1.2, c'est-à-dire que  $\mathbf{y} - \mathbf{X}\hat{\beta}$  doit former un angle droit avec  $\mathcal{S}(\mathbf{X})$ .

Puisque la matrice  $\mathbf{X}^\top\mathbf{X}$  est de plein rang, il est toujours possible de l'inverser pour résoudre les équations normales en  $\beta$  pour  $\hat{\beta}$ . On obtient la formule standard:

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \quad (1.04)$$

Même si  $\mathbf{X}$  n'est pas de plein rang, les valeurs ajustées  $\mathbf{X}\hat{\beta}$  sont définies de manière unique, parce que  $\mathbf{X}\hat{\beta}$  est tout simplement le point appartenant à  $\mathcal{S}(\mathbf{X})$  le plus près de  $\mathbf{y}$ . Regardons attentivement une fois de plus la Figure 1.2, et supposons que  $\mathbf{X}$  soit une matrice de dimension  $n \times 2$ , mais de rang égal à un seulement. Le point  $\mathbf{X}\hat{\beta}$  est toujours défini de manière unique. Cependant, si  $\beta$  est maintenant un vecteur à 2 composantes, et  $\mathcal{S}(\mathbf{X})$  est unidimensionnel, alors le vecteur  $\hat{\beta}$  n'est plus défini de manière unique. Ainsi la nécessité d'avoir une matrice  $\mathbf{X}$  de plein rang et une nécessité purement algébrique pour obtenir des estimations uniques des éléments de  $\hat{\beta}$ .

Si l'on remplaçait le terme de droite de (1.04),  $\hat{\beta}$ , par  $\mathbf{X}\hat{\beta}$ , on obtiendrait

$$\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \equiv \mathbf{P}_X\mathbf{y}. \quad (1.05)$$

Cette équation définit la matrice de dimension  $n \times n$   $\mathbf{P}_X \equiv \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ , qui **projette** le vecteur  $\mathbf{y}$  orthogonalement sur  $\mathcal{S}(\mathbf{X})$ . La matrice  $\mathbf{P}_X$  est un exemple d'une **matrice de projection orthogonale**. On associe à tout sous-espace linéaire de  $E^n$  deux matrices de ce type. L'une projette n'importe quel point de  $E^n$  sur ce sous-espace, et l'autre projette tout point de  $E^n$  sur le complément orthogonal de ce sous-espace. La matrice qui fait la projection sur  $\mathcal{S}^\perp(\mathbf{X})$  est

$$\mathbf{M}_X \equiv \mathbf{I} - \mathbf{P}_X \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top,$$

où  $\mathbf{I}$  est la matrice identité de dimension  $n \times n$ . On dit que  $\mathcal{S}(\mathbf{X})$  est le **champ** de la projection  $\mathbf{P}_X$ , alors que  $\mathcal{S}^\perp(\mathbf{X})$  est le champ de la projection  $\mathbf{M}_X$ . Notons que  $\mathbf{P}_X$  et  $\mathbf{M}_X$  sont toutes deux des matrices symétriques, et que

$$\mathbf{M}_X + \mathbf{P}_X = \mathbf{I}$$

de sorte que tout point de  $E^n$ , disons  $\mathbf{z}$ , est égal à  $\mathbf{M}_X\mathbf{z} + \mathbf{P}_X\mathbf{z}$ . Ainsi ces deux matrices définissent une **décomposition orthogonale** de  $E^n$ , parce que les

deux vecteurs  $M_X z$  et  $P_X z$  se trouvent dans deux sous-espaces orthogonaux.

A travers le livre, nous utiliserons  $P$  et  $M$  indicés par le nom d'une matrice pour désigner les matrices qui projettent respectivement, sur le sous-espace engendré par les colonnes de cette matrice et sur son complément orthogonal. Ainsi  $P_Z$  serait la matrice qui projette sur  $\mathcal{S}(Z)$ ,  $M_{X,W}$  serait la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(X, W)$ , et ainsi de suite. Ces matrices de projection ne sont toutefois d'aucune utilité pour le calcul, parce qu'elles sont de dimension  $n \times n$ , ce qui les rend trop importantes pour être manipulées par un ordinateur, sauf si la taille de l'échantillon est suffisamment restreinte. Mais elles n'en sont pas moins extrêmement utiles. Il est quelquefois pratique d'exprimer les quantités qui apparaissent en économétrie à l'aide de ces matrices, d'une part parce que les expressions qui en résultent apparaissent sous une forme compacte, et d'autre part parce que les propriétés des matrices de projection facilitent l'interprétation de ces expressions.

Dans le cas d'une régression linéaire à l'aide des régresseurs  $X$ , les matrices de projection qui sont de première importance sont  $P_X$  et  $M_X$ . Ces matrices possèdent quelques propriétés importantes, qui peuvent être décelées clairement à partir de la Figure 1.2. L'une de ces propriétés, qui est souvent très commode, est qu'elles sont **idempotentes**. Une matrice idempotente est une matrice qui multipliée par elle-même, donne la matrice de départ. Ainsi

$$P_X P_X = P_X \quad \text{et} \quad M_X M_X = M_X.$$

Ces résultats peuvent être démontrés avec un peu d'algèbre, mais l'aspect géométrique les rend évidents. Si l'on prend un point que l'on projette sur  $\mathcal{S}(X)$ , et que l'on projette sur  $\mathcal{S}(X)$  à nouveau, alors la seconde projection est sans effet, parce que le point se situe déjà sur  $\mathcal{S}(X)$ . Cela implique que  $P_X P_X z = P_X z$  pour tout vecteur  $z$ ; de sorte que  $P_X P_X = P_X$ . On pourrait tenir le même type d'argument pour  $M_X$ .

Une autre propriété importante de  $P_X$  et de  $M_X$  est que

$$P_X M_X = 0. \tag{1.06}$$

Donc  $P_X$  et  $M_X$  s'annulent l'une l'autre. Une fois encore, cela peut être démontré algébriquement par l'utilisation de  $P_X$  et de  $M_X$ , mais une telle démonstration n'est pas nécessaire. Il devrait être évident que (1.06) est vérifiée puisque  $P_X$  projette sur  $\mathcal{S}(X)$ , et  $M_X$  projette sur  $\mathcal{S}^\perp(X)$ . Le seul point qui appartienne à la fois à  $\mathcal{S}(X)$  et à  $\mathcal{S}^\perp(X)$  est l'origine, c'est-à-dire le vecteur nul. Ainsi, si nous tentons de projeter n'importe quel vecteur à la fois sur  $\mathcal{S}(X)$  et sur son complément orthogonal, nous obtenons le vecteur nul.

De fait,  $M_X$  n'annule pas seulement  $P_X$ , mais tous les points qui se trouvent sur  $\mathcal{S}(X)$ , et  $P_X$  n'annule pas seulement  $M_X$ , mais tous les points



qui se trouvent sur  $\mathcal{S}^\perp(\mathbf{X})$ . Ces propriétés peuvent à nouveau être démontrées directement par l'algèbre mais une démonstration géométrique est encore plus simple. Considérons à nouveau la Figure 1.2. Il est clair qu'en projetant n'importe quel point de  $\mathcal{S}^\perp(\mathbf{X})$  orthogonalement à  $\mathcal{S}(\mathbf{X})$ , on aboutit à l'origine (qui correspond à un vecteur dont les composantes sont des zéros), de la même façon qu'en projetant un point de  $\mathcal{S}(\mathbf{X})$  orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X})$ .

Puisque l'espace engendré par les colonnes de  $\mathbf{X}$  est invariant à une transformation linéaire non singulière des colonnes de  $\mathbf{X}$ , les matrices de projection  $\mathbf{P}_X$  et  $\mathbf{M}_X$  doivent l'être également. Cela peut être aperçu avec l'aide de l'algèbre. Considérons ce qui surviendrait si l'on postmultipliait  $\mathbf{X}$  par n'importe quelle matrice non singulière  $\mathbf{A}$  de dimension  $k \times k$ . La matrice qui projette sur l'espace engendré par  $\mathbf{XA}$  est:

$$\begin{aligned}\mathbf{P}_{\mathbf{XA}} &= \mathbf{XA}(\mathbf{A}^\top \mathbf{X}^\top \mathbf{XA})^{-1} \mathbf{A}^\top \mathbf{X}^\top \\ &= \mathbf{XAA}^{-1}(\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{A}^\top)^{-1} \mathbf{A}^\top \mathbf{X}^\top \\ &= \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{P}_X.\end{aligned}$$

Ce résultat suggère alors que le meilleur moyen de caractériser un sous-espace linéaire est de le faire grâce à la matrice qui projette orthogonalement sur lui, avec laquelle il est en correspondance biunivoque.

Si le rang de  $\mathbf{X}$  est  $k$ , alors celui de  $\mathbf{P}_X$  l'est aussi. Cela vient du fait que le champ de projection de  $\mathbf{P}_X$  est précisément  $\mathcal{S}(\mathbf{X})$ , l'espace engendré par  $\mathbf{X}$ , dont la dimension est égale à  $\rho(\mathbf{X})$ . Ainsi, bien que  $\mathbf{P}_X$  soit une matrice de dimension  $n \times n$ , son rang est généralement inférieur à  $n$ . Cet élément crucial nous permet de faire un usage de la géométrie simple plus large qu'il n'y paraissait. Puisque nous travaillons avec des vecteurs qui se situent dans un espace vectoriel à  $n$  dimensions, avec  $n$  presque toujours au moins égal à 3, il paraît impossible d'utiliser des diagrammes tels que celui que nous avons tracé dans la Figure 1.2. Mais la plupart du temps, nous serons intéressés par un sous-espace dont le nombre de dimensions est restreint, inclu dans l'espace à  $n$  dimensions dans lequel les régresseurs et la régressande sont situés. Le sous-espace dont le nombre de dimensions est restreint qui nous intéresse, sera généralement ou bien l'espace engendré par les régresseurs seulement, ou bien l'espace engendré par les régresseurs et par la régressande. Ces sous-espaces sont de dimensions  $k$  et  $k + 1$  respectivement, indépendamment de la taille  $n$  de l'échantillon. Le premier est caractérisé par la projection orthogonale  $\mathbf{P}_X$  uniquement, et le second par la projection orthogonale  $\mathbf{P}_{X,y}$ .

En examinant la figure qui serait de dimension 2, qui peut être comprise comme une projection à deux dimensions d'une image tri-dimensionnelle, les deux ou trois dimensions que l'on peut visualiser seront donc celles de  $\mathcal{S}(\mathbf{X})$  ou de  $\mathcal{S}(\mathbf{X}, \mathbf{y})$ . Ce que l'on perd en déformant les  $n$  dimensions, c'est la possibilité de dessiner des axes représentant chaque observation de l'échantillon. Pour rendre cette opération possible, il faudrait en fait se restreindre à des échantillons de deux ou trois observations. Mais cela semble

être un prix bien peu élevé à payer pour avoir la possibilité de visualiser les interprétations géométriques d'un très grand nombre de résultats algébriques en économétrie. Que de telles interprétations soient possibles découle du fait que les longueurs, les angles, les produits intérieurs et en fait, tout ce qui ne dépend pas explicitement des observations individuelles d'un ensemble de  $k$  variables linéairement indépendantes, reste inchangé lorsque l'on ignore les  $n - k$  dimensions orthogonales à l'espace engendré par les  $k$  variables. Ainsi, si par exemple deux variables sont orthogonales dans les  $n$  dimensions, elles le sont également dans l'espace à deux dimensions qu'elles engendrent.

Faisons maintenant appel aux interprétations géométriques pour établir quelques propriétés importantes du modèle de régression linéaire. Nous avons déjà vu que de (1.05),  $P_X \mathbf{y}$  est le vecteur de valeurs ajustées provenant de la régression de  $\mathbf{y}$  sur  $\mathbf{X}$ . Cela implique que  $M_X \mathbf{y} = \mathbf{y} - P_X \mathbf{y}$  est le vecteur des résidus provenant de la même régression. La propriété (1.06) nous enseigne que

$$(P_X \mathbf{y})^\top (M_X \mathbf{y}) = 0;$$

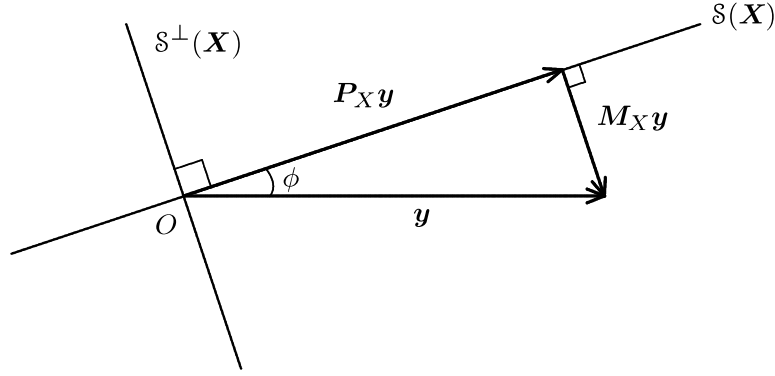
de sorte que les résidus sont orthogonaux aux valeurs ajustées. En fait, les résidus doivent être orthogonaux à tout vecteur qui se trouve dans  $\mathcal{S}(\mathbf{X})$ , ce qui inclut les régresseurs, et toute transformation linéaire de ces régresseurs. En conséquence, lorsque  $\mathbf{X}$  comprend un terme constant, ou l'équivalent d'un terme constant, les résidus doivent avoir une somme nulle.

Notons que les valeurs ajustées  $P_X \mathbf{y}$  et les résidus  $M_X \mathbf{y}$  ne dépendent de  $\mathbf{X}$  que par l'intermédiaire des matrices de projection  $P_X$  et  $M_X$ . Ainsi, ils ne dépendent que de  $\mathcal{S}(\mathbf{X})$ , et pas seulement de n'importe quelle caractéristique de  $\mathbf{X}$  qui n'affecte pas  $\mathcal{S}(\mathbf{X})$ ; en particulier, ils sont invariants à toute transformation linéaire non singulière des colonnes de  $\mathbf{X}$ . Parmi bien d'autres choses, cela implique que pour toute régression, on peut multiplier tout ou partie des régresseurs par n'importe quelle constante non nulle, et que pour toute régression qui inclut un terme constant, on peut ajouter n'importe quelle constante à tout ou partie des régresseurs, sans affecter en rien les résidus OLS ou les valeurs ajustées. A titre d'exemple, les deux régressions apparemment assez dissemblables doivent donner *exactement* les mêmes valeurs ajustées et les mêmes résidus:<sup>3</sup>

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \text{résidus};$$

$$\mathbf{y} = \alpha_1 (\mathbf{x}_1 + \mathbf{x}_2) + \alpha_2 (2\mathbf{x}_2 - \mathbf{x}_3) + \alpha_3 (3\mathbf{x}_1 - 2\mathbf{x}_2 + 5\mathbf{x}_3) + \text{résidus}.$$

<sup>3</sup> Lorsque nous écrivons une équation dont le dernier membre est “+ résidus”, nous entendons par résidus tout ce qui constitue une différence entre la régressande et la fonction de régression. Ces résidus seront les résidus OLS seulement si les paramètres de la fonction de régression sont évaluées avec les estimations OLS. Nous évitons la notation “+  $\mathbf{u}$ ” dans un tel contexte pour éviter toute référence aux propriétés statistiques des résidus.



**Figure 1.3** La décomposition orthogonale de  $\mathbf{y}$

Ces deux régressions ont le même pouvoir explicatif (c'est-à-dire les mêmes vecteurs de valeurs ajustées et de résidus) car les régresseurs engendrent exactement le même sous-espace dans les deux cas. Notons que si l'on décidait de représenter la deuxième matrice de régresseurs par  $\mathbf{X}^*$ , on verrait que  $\mathbf{X}^* = \mathbf{X}\mathbf{A}$  pour la matrice

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 3 \\ 1 & 2 & -2 \\ 0 & -1 & 5 \end{bmatrix}.$$

L'idempotence de  $\mathbf{P}_X$  et de  $\mathbf{M}_X$  rend souvent les expressions associées à la régression des moindres carrés très simples. Par exemple, évaluée en  $\hat{\beta}$ , la somme des résidus au carré (1.02) est

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) &= (\mathbf{M}_X \mathbf{y})^\top (\mathbf{M}_X \mathbf{y}) \\ &= \mathbf{y}^\top \mathbf{M}_X \mathbf{M}_X \mathbf{y} = \mathbf{y}^\top \mathbf{M}_X \mathbf{y} = \|\mathbf{M}_X \mathbf{y}\|^2. \end{aligned} \quad (1.07)$$

De manière similaire, la somme des carrés est

$$\begin{aligned} (\mathbf{X}\hat{\beta})^\top (\mathbf{X}\hat{\beta}) &= (\mathbf{P}_X \mathbf{y})^\top (\mathbf{P}_X \mathbf{y}) \\ &= \mathbf{y}^\top \mathbf{P}_X \mathbf{P}_X \mathbf{y} = \mathbf{y}^\top \mathbf{P}_X \mathbf{y} = \|\mathbf{P}_X \mathbf{y}\|^2. \end{aligned} \quad (1.08)$$

Les termes les plus à droite de (1.07) et (1.08) montrent bien que la somme des résidus au carré et la somme des carrés expliqués sont tout simplement les longueurs au carré de certains vecteurs, nommément les projections de  $\mathbf{y}$  sur les champs des deux projections  $\mathbf{M}_X$  et  $\mathbf{P}_X$ , qui sont respectivement  $\mathcal{S}^\perp(\mathbf{X})$  et  $\mathcal{S}(\mathbf{X})$ .

Cela est montré sur la Figure 1.3, qui correspond à la Figure 1.2 redessinée et avec de nouvelles étiquettes. Bien qu'elle soit bi-dimensionnelle, c'est une figure parfaitement générale. L'espace bi-dimensionnel représenté

sur la figure est celui engendré par la régressande  $\mathbf{y}$  et par le vecteur de valeurs ajustées  $\mathbf{P}_X \mathbf{y}$ . Ces deux vecteurs forment, comme on peut le voir, deux côtés d'un triangle rectangle. La distance entre  $\mathbf{y}$  et  $\mathbf{P}_X \mathbf{y}$  est  $\|\mathbf{M}_X \mathbf{y}\|$ , la distance entre l'origine et  $\mathbf{P}_X \mathbf{y}$  est  $\|\mathbf{P}_X \mathbf{y}\|$ , et la distance entre l'origine et  $\mathbf{y}$  est bien évidemment  $\|\mathbf{y}\|$ . En appliquant le Théorème de Pythagore, on conclue immédiatement que,<sup>4</sup>

$$\|\mathbf{y}\|^2 = \|\mathbf{P}_X \mathbf{y}\|^2 + \|\mathbf{M}_X \mathbf{y}\|^2$$

de sorte que la **somme des carrés totaux**, ou **TSS**, de la régressande est égale à la somme des carrés expliqués plus la somme des résidus au carré. Ce résultat dépend de manière cruciale du fait que  $\mathbf{M}_X \mathbf{y}$  est orthogonal à  $\mathcal{S}(\mathbf{X})$ , puisqu'autrement nous ne pourrions pas obtenir un angle droit dans le triangle, et nous ne pourrions pas appliquer le Théorème de Pythagore.

Le fait que la variation de la régressande puisse être divisée en deux parties, l'une “expliquée” par les régresseurs, et l'autre qui reste sans explication, suggère une mesure naturelle de la qualité de l'ajustement. Cette mesure, appelée formellement **coefficient de détermination**, mais à laquelle on fait universellement référence en tant que  $R^2$ , a plusieurs variantes. Sa version la plus simple (mais pas celle la plus répandue) est le  $R^2$  **non centré**:

$$R_u^2 = \frac{\|\mathbf{P}_X \mathbf{y}\|^2}{\|\mathbf{y}\|^2} = 1 - \frac{\|\mathbf{M}_X \mathbf{y}\|^2}{\|\mathbf{y}\|^2}.$$

Il est clair que le  $R_u^2$  ne possède aucune unité et qu'il doit prendre des valeurs entre zéro et un. De la Figure 1.3, il est facile de voir que  $R_u^2$  a une interprétation géométrique simple. Le cosinus de l'angle formé par le vecteur  $\mathbf{y}$  et  $\mathbf{P}_X \mathbf{y}$ , que nous avons noté  $\phi$  sur la figure, est

$$\cos \phi = \frac{\|\mathbf{P}_X \mathbf{y}\|}{\|\mathbf{y}\|}.$$

C'est le coefficient de corrélation (non centré) entre  $\mathbf{y}$  et  $\mathbf{P}_X \mathbf{y}$ . Par conséquent, nous voyons que

$$R_u^2 = \cos^2 \phi.$$

Lorsque  $\mathbf{y}$  se trouve véritablement dans  $\mathcal{S}(\mathbf{X})$ , l'angle formé par  $\mathbf{P}_X \mathbf{y}$  et  $\mathbf{y}$  doit être nul, puisque ce sont les mêmes vecteurs, et donc  $R_u^2$  sera égal à l'unité. Dans le cas extrême opposé, si  $\mathbf{y}$  appartient à  $\mathcal{S}^\perp(\mathbf{X})$ , l'angle entre  $\mathbf{P}_X \mathbf{y}$  et  $\mathbf{y}$  sera de  $90^\circ$ , et le  $R_u^2$  sera alors nul.

Tout ce qui modifie l'angle  $\phi$  de la Figure 1.3 modifiera le  $R^2$  non centré. En particulier, il est évident qu'en ajoutant une constante à  $\mathbf{y}$  on modifiera

<sup>4</sup> **Le Théorème de Pythagore** enseigne que pour un triangle rectangle, le carré de l'hypothénuse égale la somme des carrés des côtés adjacents.

cet angle, et cela est vrai même si  $\mathbf{X}$  inclut un terme constant. Si  $R^2$  doit être utilisé en tant que mesure de la qualité d'ajustement, il paraît indésirable qu'il change lors d'une opération aussi simple que l'addition d'un terme constant à la régressande. Il est facile de modifier le  $R^2$  pour contourner ce problème. La version modifiée est connue sous le nom de  $R^2$  **centré**, et nous le noterons  $R_c^2$ . On le définit par

$$R_c^2 \equiv 1 - \frac{\|\mathbf{M}_X \mathbf{y}\|^2}{\|\mathbf{M}_\iota \mathbf{y}\|^2}, \quad (1.09)$$

où

$$\mathbf{M}_\iota \equiv \mathbf{I} - \iota(\iota^\top \iota)^{-1} \iota^\top = \mathbf{I} - n^{-1} \iota \iota^\top$$

est la matrice qui projette en dehors de l'espace engendré par le vecteur de constantes  $\iota$ , qui est un vecteur dont les composantes sont égales à l'unité. Lorsqu'un vecteur est multiplié par  $\mathbf{M}_\iota$ , le résultat est un vecteur de déviations par rapport à la moyenne. Ainsi, ce que mesure le  $R^2$  centré correspond à la proportion de la somme des carrés totaux de la régressande *autour de sa moyenne* qui est expliquée par les régresseurs.

Une expression alternative de  $R_c^2$  est:

$$\frac{\|\mathbf{P}_X \mathbf{M}_\iota \mathbf{y}\|^2}{\|\mathbf{M}_\iota \mathbf{y}\|^2}, \quad (1.10)$$

qui est identique à (1.09) si  $\mathbf{P}_X \iota = \iota$ , ce qui signifie que  $\mathcal{S}(\mathbf{X})$  doit contenir le vecteur  $\iota$  (de sorte que soit une colonne de  $\mathbf{X}$  doit être une constante, soit une combinaison linéaire des colonnes de  $\mathbf{X}$  doit être égale à une constante). Dans ce cas, l'égalité doit être vérifiée parce que

$$\mathbf{M}_X \mathbf{M}_\iota \mathbf{y} = \mathbf{M}_X (\mathbf{I} - \mathbf{P}_\iota) \mathbf{y} = \mathbf{M}_X \mathbf{y},$$

la seconde égalité provenant du fait que  $\mathbf{M}_X$  annule  $\mathbf{P}_\iota$  lorsque  $\iota$  appartient à  $\mathcal{S}(\mathbf{X})$ . Lorsque cela n'est pas le cas et que (1.10) n'est plus vérifiée, il n'y a aucune garantie que  $R_c^2$  soit positif. Après tout, il y aura de nombreux cas où la régressande  $\mathbf{y}$  sera mieux expliquée par un terme constant que par un ensemble de régresseurs dans lequel il n'y a pas de terme constant. Clairement, si (1.10) est vérifiée,  $R_c^2$  doit appartenir à  $[0;1]$  puisque (1.10) est alors simplement le  $R^2$  non centré pour une régression de  $\mathbf{M}_\iota \mathbf{y}$  sur  $\mathbf{X}$ .

L'usage du  $R^2$  centré lorsque  $\mathbf{X}$  n'inclut pas de terme constant ou d'équivalent est ainsi source de difficultés. Quelques programmes pour les statistiques ou l'économétrie n'imprimeront tout simplement pas le  $R^2$  dans ces circonstances; d'autres donnent le  $R_u^2$  (sans toujours avertir l'utilisateur qu'ils calculent cet indice); certains affichent le  $R_c^2$ , tel que nous l'avons défini dans (1.09) qui est soit positif, soit négatif; et certains affichent d'autres quantités, qui seraient égales à  $R_c^2$  si  $\mathbf{X}$  contenait un terme constant, mais qui ne le sont pas dans le cas contraire. Utilisateurs de logiciels de statistique, soyez méfiants!

Notons que le  $R^2$  est un indice intéressant uniquement parce que nous avons fait usage des moindres carrés pour estimer  $\hat{\beta}$ . Si nous choisissons une estimation de  $\beta$ , disons  $\tilde{\beta}$ , d'une autre façon, de sorte que le triangle de la Figure 1.3 ne comporte plus d'angle droit, nous concluons que les deux définitions du  $R^2$ , (1.09) et (1.10), ne sont plus identiques:

$$1 - \frac{\|y - X\tilde{\beta}\|^2}{\|y\|^2} \neq \frac{\|X\tilde{\beta}\|^2}{\|y\|^2}.$$

Si nous choisissons de définir  $R^2$  en fonction des résidus, en utilisant la première expression, nous ne pourrions garantir qu'il serait positif; et si nous choisissons de le définir en fonction des valeurs ajustées, nous ne pourrions garantir qu'il soit inférieur à l'unité. Ainsi, lorsque l'on utilise autre chose que les moindres carrés pour estimer une régression, l'on devrait ignorer ce que l'on a appelé  $R^2$ , ou bien, être certain de savoir ce qui s'y rapporte exactement dans le calcul de l'ordinateur.

### 1.3 RESTRICTIONS ET REPARAMÉTRISATIONS

Nous avons insisté sur le fait que  $\mathcal{S}(X)$  est invariant à une transformation linéaire non singulière des colonnes de  $X$ . Cela implique que l'on peut toujours **reparamétriser** n'importe quelle régression pour la rendre plus commode, sans changer en rien la capacité d'explication de la régressande. Supposons que l'on veuille opérer la régression

$$y = X\beta + \text{résidus} \quad (1.11)$$

et comparer les résultats avec ceux d'une autre régression, où  $\beta$  est soumis aux  $r(\leq k)$  restrictions linéairement indépendantes

$$R\beta = r, \quad (1.12)$$

où  $R$  est une matrice de dimension  $r \times k$  et de rang  $r$ , où  $r$  est un vecteur à  $r$  composantes. Bien que cela ne pose pas de difficulté avec les **moindres carrés contraints**, il est parfois plus aisé de reparamétriser la régression pour faire en sorte que les restrictions soient nulles. Alors la régression contrainte peut être estimée de la manière usuelle par **OLS**. La reparamétrisation peut s'effectuer comme suit.

Premièrement, il faut réorganiser les colonnes de  $X$  de façon que les restrictions (1.12) puissent s'écrire

$$R_1\beta_1 + R_2\beta_2 = r, \quad (1.13)$$

où  $\mathbf{R} \equiv [\mathbf{R}_1 \ \mathbf{R}_2]$  et  $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}_1 \ ; \ \boldsymbol{\beta}_2]$ <sup>5</sup>  $\mathbf{R}_1$  étant une matrice de dimension  $r \times r$  et non singulière,  $\mathbf{R}_2$  étant de dimension  $r \times (k - r)$ . Résoudre les équations (1.13) pour  $\boldsymbol{\beta}_1$  entraîne

$$\boldsymbol{\beta}_1 = \mathbf{R}_1^{-1}\mathbf{r} - \mathbf{R}_1^{-1}\mathbf{R}_2\boldsymbol{\beta}_2.$$

Ainsi la régression originelle (1.11), avec les restrictions que l'on a imposées peut s'écrire:

$$\mathbf{y} = \mathbf{X}_1(\mathbf{R}_1^{-1}\mathbf{r} - \mathbf{R}_1^{-1}\mathbf{R}_2\boldsymbol{\beta}_2) + \mathbf{X}_2\boldsymbol{\beta}_2 + \text{résidus}.$$

Ce qui équivaut à

$$\mathbf{y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r} = (\mathbf{X}_2 - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{R}_2)\boldsymbol{\beta}_2 + \text{résidus},$$

qui, en définissant  $\mathbf{y}^*$  comme étant égal à  $\mathbf{y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$ , et  $\mathbf{Z}_2$  égal à  $\mathbf{X}_2 - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{R}_2$ , peut être exprimée de manière plus ramassée comme

$$\mathbf{y}^* = \mathbf{Z}_2\boldsymbol{\beta}_2 + \text{résidus}. \quad (1.14)$$

Ceci correspond à la version contrainte de la régression (1.11). Pour retrouver une régression équivalente à la régression de départ, il nous faut ajouter  $r$  régresseurs, qui, avec  $\mathbf{Z}_2$ , engendreront le même espace que  $\mathbf{X}$ . Il y a une infinité de façons de le faire. Notons la nouvelle fonction de régression  $\mathbf{Z}\boldsymbol{\gamma}$ . Nous avons déjà défini  $\mathbf{Z}_2$ , et  $\mathbf{Z}_1$  peut être n'importe quelle matrice comportant  $r$  colonnes qui, conjointement à  $\mathbf{Z}_2$ , engendre  $\mathcal{S}(\mathbf{X})$ . Il suit que nous devons avoir  $\boldsymbol{\gamma}_2 = \boldsymbol{\beta}_2$ . La nouvelle régression est ainsi

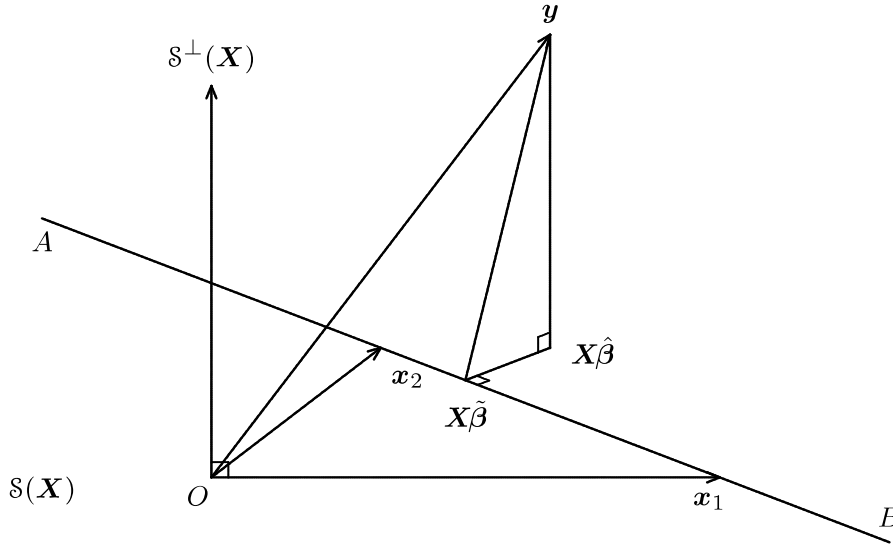
$$\mathbf{y}^* = \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{Z}_2\boldsymbol{\gamma}_2 + \text{résidus}. \quad (1.15)$$

Il devrait être évident que de la manière dont nous avons construit  $\mathbf{Z}_2$ ,  $\mathcal{S}(\mathbf{X}_1, \mathbf{Z}_2) = \mathcal{S}(\mathbf{X})$ , et qu'un choix possible pour  $\mathbf{Z}_1$  soit tout simplement  $\mathbf{X}_1$ . Le fait que la régressande ait été également modifiée ne peut pas avoir d'effet sur les résidus de la régression parce que  $\mathbf{y}^* = \mathbf{y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$ , et que le vecteur  $\mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$  se situe dans  $\mathcal{S}(\mathbf{X})$ .

A titre d'exemple de la procédure que nous venons de décrire, considérons la régression:

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \text{résidus}, \quad (1.16)$$

<sup>5</sup> La notation  $[\boldsymbol{\beta}_1 \ ; \ \boldsymbol{\beta}_2]$  signifie que  $\boldsymbol{\beta}_1$  et  $\boldsymbol{\beta}_2$  sont deux vecteurs colonne empilés de manière à former un autre vecteur colonne, dans ce cas  $\boldsymbol{\beta}$ . Une notation plus commune serait  $[\boldsymbol{\beta}_1^\top \ \boldsymbol{\beta}_2^\top]^\top$ , mais elle est à l'évidence assez gênante. Nous introduisons la notation précédente parce que l'on a souvent besoin d'empiler des vecteurs colonne et nous voyons que cette nouvelle notation est assez innovatrice pour mériter d'être introduite.



**Figure 1.4** Estimation contrainte et non contrainte

qui doit être estimée sous la contrainte  $\beta_1 + \beta_2 = 1$ . La régression contrainte équivalente à (1.14) est donc:

$$y - x_1 = \beta_2(x_2 - x_1) + \text{résidus},$$

et la régression non contrainte dans la nouvelle paramétrisation équivalente à (1.15) est:

$$y^* = \gamma_1 z_1 + \gamma_2 z_2 + \text{résidus}, \quad (1.17)$$

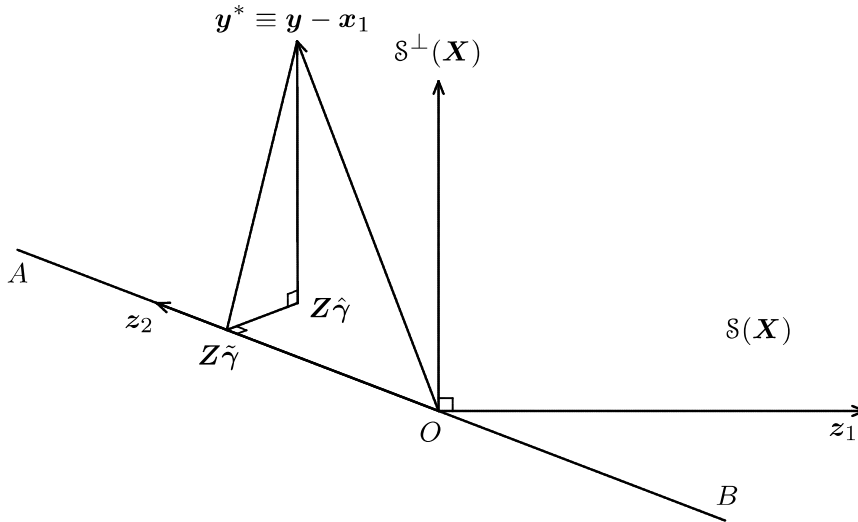
dans laquelle  $y^* \equiv y - x_1$ ,  $z_1 \equiv x_1$ ,  $z_2 \equiv x_2 - x_1$ , et  $\gamma_2 \equiv \beta_2$ . La contrainte de nullité de  $\gamma_1$  est équivalente à la restriction originelle  $\beta_1 + \beta_2 = 1$ .

Cet exemple est illustré par la Figure 1.4 qui est une figure tri-dimensionnelle. La figure rassemble l'espace engendré par les deux régresseurs  $x_1$  et  $x_2$ , qui correspond au plan noté  $S(X)$  sur la figure et l'espace engendré par la régressande  $y$ , qui doit être vu comme étant au dessus de  $S(X)$ . L'estimation OLS non contrainte correspond à la projection orthogonale de  $y$  sur la surface du plan, au point  $X\hat{\beta}$ . La restriction  $\beta_1 + \beta_2 = 1$  implique que les valeurs ajustées se situent sur la ligne  $AB$  sur la figure, et lorsque l'on projette  $y$  orthogonalement sur cette ligne, on obtient le point  $X\tilde{\beta}$ , qui représente le vecteur de valeurs ajustées de la régression contrainte.

La reparamétrisation qui fait passer de (1.16) à (1.17) ne modifie en rien  $S(X)$ , mais puisque la régressande est désormais  $y - x_1$  au lieu de  $y$ , la localisation de la régressande est modifiée par rapport à l'origine. La Figure 1.5 est sensiblement la même que la Figure 1.4, excepté que tout a été étiqueté avec les termes de la nouvelle paramétrisation.

En fait, la Figure 1.5 résulte de la Figure 1.4 en déplaçant l'origine à l'extrémité de la flèche représentant la variable  $x_1$ . Le vecteur  $z_1$  est ainsi

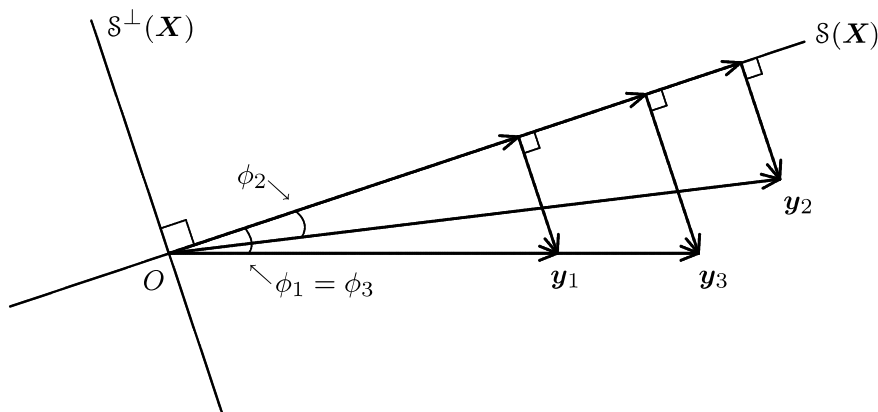




**Figure 1.5** Estimation après reparamétrisation

l'ancien vecteur  $x_1$  translaté, de manière à partir de la nouvelle origine. Le second nouveau régresseur,  $z_2$ , se situe le long de la ligne  $AB$ , qui traverse l'origine sur la Figure 1.5. Cela doit rendre évident le fait que la restriction signifie que  $\gamma_1$  doit être nul, de manière à ce que le vecteur de valeurs ajustées se situe lui aussi le long de la ligne  $AB$ .

Nous avons vu que les résidus de la régression reparamétrisée (1.15) seront exactement les mêmes que les résidus de (1.11), et ce fait est évident d'après les Figures 1.4 et 1.5. Cela ne sera pas exact, que ce soit pour le  $R^2$  centré ou pour le  $R^2$  non centré, puisque la somme des carrés totaux dépend de la façon dont on exprime la régressande. C'est une raison supplémentaire pour être prudent vis à vis de toutes les versions du  $R^2$ : des régressions équivalentes peuvent avoir différentes valeurs du  $R^2$ . Pour en avoir la preuve géométrique, considérons la Figure 1.6, qui est très semblable à la Figure 1.3, sauf que nous avons maintenant trois régressandes,  $y_1$ ,  $y_2$ , et  $y_3$ , qui sont tracées. La deuxième régressande  $y_2$  a été obtenue à partir de  $y_1$  en déplaçant celle-ci vers le nord-est, parallèlement à  $S(X)$ , de sorte que  $M_X y_2 = M_X y_1$ . Notons que  $\phi_1$  est différent de  $\phi_2$ , de manière à rendre des  $R^2$  différents. D'autre part,  $y_3$  a été obtenue en prolongeant  $y_1$ , et en maintenant  $\phi$  constant. En conséquence,  $M_X y_3$  sera plus important que  $M_X y_1$ , mais les deux régressions auront le même  $R^2$  (non centré). Si l'on interprétait  $S(X)$  dans la figure comme  $S(M_t X)$  et  $y$  comme  $M_t y$ , alors le cosinus au carré de  $\phi$  ( $\cos^2 \phi$ ) serait le  $R^2$  centré, au lieu d'être le  $R^2$  non centré.



**Figure 1.6** Effets sur le  $R^2$  des différentes régressandes

## 1.4 LE THÉORÈME FRISCH-WAUGH-LOVELL

Nous allons à présent discuter d'une propriété extrêmement importante et utile des estimations par moindres carrés, qui, bien que largement diffusée et connue, n'est pas aussi largement reconnue qu'elle devrait l'être. Nous nous y référerons en tant que **Théorème Frisch-Waugh-Lovell**, ou **Théorème FWL**, d'après Frisch et Waugh (1933) et Lovell (1963), puisque ces articles semblent l'avoir introduit, puis réintroduit dans l'outillage des économètres. Le théorème est beaucoup plus général cependant, et d'un usage plus général qu'une lecture négligée des articles le laisserait supposer. Parmi d'autres choses, il élimine presque entièrement la nécessité d'inverser les matrices partitionnées lorsque l'on dérive certains résultats standards des moindres carrés ordinaires (et non linéaires).

Le Théorème FWL s'applique à toute régression où il y a deux ou plusieurs régresseurs, que l'on peut partitionner logiquement en deux groupes. La régression s'écrit ainsi:

$$y = X_1\beta_1 + X_2\beta_2 + \text{résidus}, \quad (1.18)$$

où  $X_1$  est de dimension  $n \times k_1$ ,  $X_2$  de dimension  $n \times k_2$ , avec  $X \equiv [X_1 \ X_2]$  et  $k = k_1 + k_2$ . Par exemple,  $X_1$  pourrait être un groupe de variables saisonnières muettes, ou des variables de tendance, et  $X_2$  un groupe de véritables variables économiques. Ceci était en fait le genre de situation traitée par Frisch et Waugh (1933) et Lovell (1963). Un autre exemple serait de regrouper dans  $X_1$  les régresseurs dont nous désirons tester la signification jointe, et dans  $X_2$  d'autres régresseurs que nous ne testerons pas. Ou encore,  $X_1$  regrouperait les régresseurs dont nous savons qu'ils sont orthogonaux à la régressande, et  $X_2$  les régresseurs qui ne le sont pas, une situation qui arrive fréquemment lorsque l'on désire tester les modèles de régression non linéaire; consulter le Chapitre 6.

Maintenant, considérons une autre régression

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \boldsymbol{\beta}_2 + \text{résidus}, \quad (1.19)$$

où  $\mathbf{M}_1$  est une matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}_1)$ . Dans (1.19), nous avons tout d'abord régressé  $\mathbf{y}$  et chacune des  $k_2$  colonnes de  $\mathbf{X}_2$  sur  $\mathbf{X}_1$ , puis régressé le vecteur de résidus  $\mathbf{M}_1 \mathbf{y}$  sur la matrice  $\mathbf{M}_1 \mathbf{X}_2$  de résidus de dimension  $n \times k_2$ . Le Théorème FWL nous affirme que les résidus des régressions (1.18) et (1.19), ainsi que les estimations OLS de  $\boldsymbol{\beta}_2$  provenant de ces deux régressions, seront numériquement identiques. Géométriquement, dans la régression (1.18), on projette  $\mathbf{y}$  directement sur  $\mathcal{S}(\mathbf{X}) \equiv \mathcal{S}(\mathbf{X}_1, \mathbf{X}_2)$ , alors que dans la régression (1.19) on projette tout d'abord  $\mathbf{y}$  et toutes les colonnes de  $\mathbf{X}_2$  en dehors de  $\mathcal{S}(\mathbf{X}_1)$ , puis on projette les résidus  $\mathbf{M}_1 \mathbf{y}$  sur l'espace engendré par la matrice des résidus,  $\mathcal{S}(\mathbf{M}_1 \mathbf{X}_2)$ . Alors le Théorème FWL nous indique que ces procédures apparemment dissemblables entraînent pourtant les mêmes résultats.

On peut démontrer le Théorème FWL de plusieurs façons différentes. Une démonstration assez connue est basée sur l'algèbre des matrices partitionnées. Tout d'abord, observons que les estimations de  $\boldsymbol{\beta}_2$  provenant de (1.19) sont

$$(\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}. \quad (1.20)$$

Cette expression simple dont nous ferons un usage fréquent, provient directement de la substitution de  $\mathbf{X}$  par  $\mathbf{M}_1 \mathbf{X}_2$ , et de  $\mathbf{y}$  par  $\mathbf{M}_1 \mathbf{y}$ , dans l'expression (1.04) pour le vecteur des estimations OLS. La démonstration algébrique ferait ensuite appel aux résultats sur l'inversion des matrices partitionnées (consulter l'Annexe A) pour montrer que l'estimation  $\hat{\boldsymbol{\beta}}_2$  par OLS de (1.18) est identique à (1.20), et finirait par montrer que les deux groupes de résidus sont aussi identiques. Nous laissons cette démonstration pour exercice, et poursuivons dans un premier temps par une discussion plus détaillée de la géométrie de la situation.

Soit  $\hat{\boldsymbol{\beta}} \equiv [\hat{\boldsymbol{\beta}}_1 \ ; \ \hat{\boldsymbol{\beta}}_2]$  les estimations OLS de (1.18). Alors

$$\mathbf{y} = \mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y} = \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2 + \mathbf{M}_X \mathbf{y}. \quad (1.21)$$

En multipliant  $\mathbf{y}$  et le terme de la droite de (1.21), qui est égal à  $\mathbf{y}$ , par  $\mathbf{X}_2^\top \mathbf{M}_1$ , où  $\mathbf{M}_1 \equiv \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top$ , on obtient

$$\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} = \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 \hat{\boldsymbol{\beta}}_2. \quad (1.22)$$

Le premier terme du membre de droite de (1.21) a disparu parce que  $\mathbf{M}_1$  annule  $\mathbf{X}_1$ . Le dernier terme a disparu parce que  $\mathbf{M}_1 \mathbf{X}_2 = \mathbf{X}_2 - \mathbf{P}_1 \mathbf{X}_2$  appartient à  $\mathcal{S}(\mathbf{X})$ , de sorte que

$$\mathbf{M}_X \mathbf{M}_1 \mathbf{X}_2 = \mathbf{0}. \quad (1.23)$$

En résolvant (1.22) pour  $\hat{\beta}_2$ , nous observons immédiatement que

$$\hat{\beta}_2 = (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y},$$

qui correspond à (1.20). Cela apporte la démonstration à la seconde partie du théorème.

Si l'on avait multiplié (1.21) par  $\mathbf{M}_1$  au lieu de  $\mathbf{X}_2^\top \mathbf{M}_1$ , nous aurions obtenu

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2 + \mathbf{M}_X \mathbf{y}. \quad (1.24)$$

A présent, la régressande est la régressande de la régression (1.19). Parce que  $\hat{\beta}_2$  est l'estimation de  $\beta_2$  de (1.19), le premier terme du membre de droite de (1.24) est le vecteur de valeurs ajustées de cette régression. Ainsi le second terme doit correspondre au vecteur de résidus de la régression (1.19). Mais  $\mathbf{M}_X \mathbf{y}$  est aussi le vecteur de résidus de la régression (1.18), de manière à démontrer la première partie du théorème.

Considérons donc maintenant l'aspect géométrique dans le détail. Parce que  $\mathcal{S}(\mathbf{X}_1)$  et  $\mathcal{S}(\mathbf{X}_2)$  ne sont pas en général des espaces mutuellement orthogonaux, les deux premiers termes du membre de droite de (1.21),  $\mathbf{X}_1 \hat{\beta}_1$  et  $\mathbf{X}_2 \hat{\beta}_2$ , ne sont donc pas non plus orthogonaux en général. Si l'on décompose  $\mathbf{X}_2 \hat{\beta}_2$ , il vient:

$$\mathbf{X}_2 \hat{\beta}_2 = \mathbf{P}_1 \mathbf{X}_2 \hat{\beta}_2 + \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2 \quad (1.25)$$

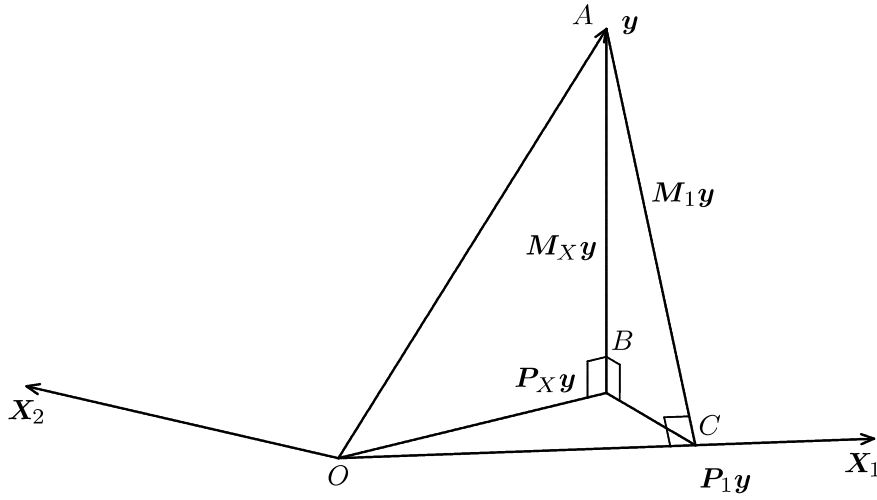
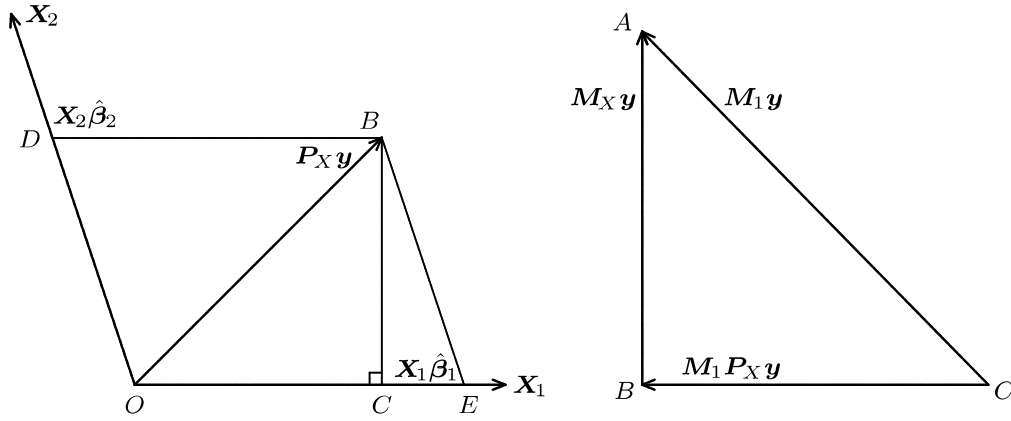
et si l'on regroupe les termes de (1.21), on arrive au résultat:

$$\begin{aligned} \mathbf{y} &= (\mathbf{X}_1 \hat{\beta}_1 + \mathbf{P}_1 \mathbf{X}_2 \hat{\beta}_2) + \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2 + \mathbf{M}_X \mathbf{y} \\ &= \mathbf{P}_1 (\mathbf{X}_1 \hat{\beta}_1 + \mathbf{X}_2 \hat{\beta}_2) + \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2 + \mathbf{M}_X \mathbf{y} \\ &= \mathbf{P}_1 \mathbf{y} + \mathbf{M}_1 \mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y}, \end{aligned} \quad (1.26)$$

parce que  $(\mathbf{X}_1 \hat{\beta}_1 + \mathbf{P}_1 \mathbf{X}_2 \hat{\beta}_2)$  est évidemment  $\mathbf{P}_1 \mathbf{P}_X \mathbf{y} = \mathbf{P}_1 \mathbf{y}$ , alors que  $\mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2$  est égal à  $\mathbf{M}_1 (\mathbf{X}_1 \hat{\beta}_1 + \mathbf{X}_2 \hat{\beta}_2)$ , qui peut être écrit par  $\mathbf{M}_1 \mathbf{P}_X \mathbf{y}$ . La dernière expression de (1.26) met en évidence le fait que  $\mathbf{y}$  est la somme de trois termes mutuellement orthogonaux. Le résultat (1.26) implique que

$$\mathbf{M}_1 \mathbf{y} = \mathbf{y} - \mathbf{P}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y}. \quad (1.27)$$

Considérons désormais la Figure 1.7 dans laquelle ceci est reporté pour le cas le plus simple pour lequel le théorème s'applique, c'est-à-dire le cas où  $k_1 = k_2 = 1$ . Sur le schéma (a) de la figure, qui représente un graphe tri-dimensionnel, le vecteur  $\mathbf{y}$  est dessiné avec sa projection  $\mathbf{P}_X \mathbf{y}$  sur la surface engendrée par les deux régresseurs  $\mathbf{X}_1$  et  $\mathbf{X}_2$ , correspondant au plan horizontal, et avec sa projection complémentaire (verticale)  $\mathbf{M}_X \mathbf{y}$ . Nous avons aussi représenté la projection  $\mathbf{P}_1 \mathbf{y}$  de  $\mathbf{y}$  sur la direction du premier régresseur  $\mathbf{X}_1$  et son complément  $\mathbf{M}_1 \mathbf{y}$ . Observons que le triangle  $ABC$ , formé

(a) Deux projections de  $y$ (b) L'espace  $\mathcal{S}(\mathbf{X})$  engendré par les régresseurs(c) La décomposition de  $M_1 y$ **Figure 1.7** Le Théorème de Frisch-Waugh-Lovell

par le point  $y$ , le point  $P_X y$ , et par le point  $P_1 y$ , est un triangle rectangle dans le plan vertical, perpendiculaire à la direction de  $X_1$ .

Le schéma bi-dimensionnel (b) représente le plan horizontal qui est le champ  $\mathcal{S}(\mathbf{X})$  de la projection  $P_X$ . La décomposition de  $P_X y$  en  $X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2$  est illustrée via le parallélogramme  $ODBE$ , où les côtés  $OE$  et  $DB$  correspondent à  $X_1 \hat{\beta}_1$ , et les côtés  $OD$  et  $EB$  représentent  $X_2 \hat{\beta}_2$ . Le triangle  $EBC$  illustre la décomposition (1.25) de  $X_2 \hat{\beta}_2$ , en la somme de  $P_1 X_2 \hat{\beta}_2$ , représenté par  $EC$ , et de  $M_1 X_2 \hat{\beta}_2 = M_1 P_X y$ , représenté par  $CB$ . Notons que ces deux vecteurs se trouvent dans  $\mathcal{S}(\mathbf{X})$ ; le fait que le second se situe effectivement dans  $\mathcal{S}(\mathbf{X})$  provient de l'équation (1.23). Enfin, le schéma (c) montre en deux dimensions le triangle  $ABC$ . Cela représente la décomposition de  $M_1 y$ ,

donnée par (1.27), en  $\mathbf{M}_1 \mathbf{P}_X \mathbf{y}$ , correspondant à  $CB$ , et  $\mathbf{M}_X \mathbf{y}$ , correspondant à  $BA$ .

Le dernier schéma peut désormais être utilisé pour illustrer le Théorème FWL. L'élément qu'il faut comprendre est que la décomposition orthogonale (1.27) de  $\mathbf{M}_1 \mathbf{y}$ , en la somme de  $\mathbf{M}_X \mathbf{y}$  et  $\mathbf{M}_1 \mathbf{P}_X \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2$ , est la décomposition orthogonale qui est opérée par la régression de  $\mathbf{M}_1 \mathbf{y}$ , le vecteur décomposé, sur les colonnes de  $\mathbf{M}_1 \mathbf{X}_2$ , ou pour aller vite, la régression (1.19). Cela doit être clair géométriquement; de manière algébrique, il découle premièrement, du fait que le terme  $\mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2$  est évidemment une décomposition linéaire des colonnes de  $\mathbf{M}_1 \mathbf{X}_2$  et deuxièmement, du fait que l'autre terme,  $\mathbf{M}_X \mathbf{y}$ , est orthogonal à toutes ces colonnes. Ce second élément découle de la relation  $\mathbf{M}_X \mathbf{M}_1 = \mathbf{M}_X$ , qui est vérifiée parce que  $\mathcal{S}^\perp(\mathbf{X})$  est un sous-espace de  $\mathcal{S}^\perp(\mathbf{X}_1)$ , et pour cela, comme nous l'avons vu dans l'équation (1.23),

$$\mathbf{M}_X \mathbf{M}_1 \mathbf{X}_2 = \mathbf{M}_X \mathbf{X}_2 = \mathbf{0},$$

de sorte que  $\mathbf{y}^\top \mathbf{M}_X \mathbf{M}_1 \mathbf{X}_2 = \mathbf{0}$ . Ainsi, nous avons montré que

$$\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \hat{\beta}_2 \quad \text{et} \quad (1.28)$$

$$\mathbf{M}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y} = \mathbf{M}_X \mathbf{y}. \quad (1.29)$$

Une partie du Théorème FWL établit que les régressions (1.18) et (1.19) ont les mêmes résidus. Ces résidus communs sont constitués par le vecteur  $\mathbf{M}_X \mathbf{y}$ , comme cela est clair de (1.21) pour la régression (1.18), et de (1.29) pour la régression (1.19). L'autre volet du théorème établit que les estimations de  $\hat{\beta}_2$  sont les mêmes pour les deux régressions. Cela est maintenant évident à partir de (1.28), dans laquelle le  $\hat{\beta}_2$  de la régression (1.18) entraîne un vecteur de valeurs ajustées  $\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y}$  pour la régression (1.19).

Nous rencontrerons nombre d'applications du Théorème FWL à travers tout le livre. Un exemple simple est l'utilisation de variables muettes pour l'ajustement saisonnier, qui était l'application originelle de Lovell (1963). Plusieurs séries temporelles économiques qui sont collectées mensuellement ou trimestriellement, manifestent des variations saisonnières symétriques. Une manière de modéliser ce phénomène est d'adjoindre un ensemble de variables saisonnières muettes à la régression. Par exemple, supposons que les données sont trimestrielles, que les variables muettes de saisonnalité  $\mathbf{D}_1$ ,  $\mathbf{D}_2$  et  $\mathbf{D}_3$  sont définies ainsi:

$$\mathbf{D}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ -1 \\ \vdots \end{bmatrix} \quad \mathbf{D}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ -1 \\ \vdots \end{bmatrix} \quad \mathbf{D}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 1 \\ -1 \\ \vdots \end{bmatrix}.$$

Notons que ces variables muettes ont été définies de telle manière que leur somme est nulle sur chaque année. Considérons les régressions

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\gamma} + \text{résidus} \quad \text{et} \quad (1.30)$$

$$\mathbf{M}_D \mathbf{y} = \mathbf{M}_D \mathbf{X} \boldsymbol{\beta} + \text{résidus}, \quad (1.31)$$

où  $\mathbf{D} \equiv [\mathbf{D}_1 \quad \mathbf{D}_2 \quad \mathbf{D}_3]$  et  $\mathbf{M}_D$  est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{D})$ . Dans (1.30) nous incluons des variables saisonnières muettes dans une régression où les données ne sont pas ajustées. Dans (1.31) on ajuste les données pour chaque saison en les régressant sur les variables muettes et ensuite, on exécute la régression sur les données “ajustées par saison”. Le Théorème FWL implique que ces deux procédures entraînent des estimations de  $\boldsymbol{\beta}$  identiques.

La régressande  $\mathbf{M}_D \mathbf{y}$  et les régresseurs  $\mathbf{M}_D \mathbf{X}$  qui sont mentionnés dans (1.31), peuvent être considérés comme des versions ajustées par saison de  $\mathbf{y}$  et de  $\mathbf{X}$ , parce que toutes les variations dans  $\mathbf{y}$  et dans  $\mathbf{X}$  qui peuvent être attribuées à des différences systématiques dans la moyenne trimestrielle ont été éliminées de  $\mathbf{M}_D \mathbf{y}$  et de  $\mathbf{M}_D \mathbf{X}$ . Ainsi, l'équivalence entre (1.30) et (1.31) est souvent utilisée pour justifier l'idée selon laquelle il importe peu d'utiliser des données brutes ou des données ajustées par saison lorsque l'on estime un modèle de régression comprenant des données temporelles. Malheureusement, une telle conclusion n'est pas toujours fiable. Les procédures officielles d'ajustement saisonnier ne sont pas toujours basées sur la régression, de sorte que l'utilisation de données ajustées par saison n'est pas équivalente à l'utilisation des résidus de la régression sur un ensemble de variables muettes. De plus, si (1.30) n'est pas une description raisonnable (et elle n'en serait pas une si, par exemple, le modèle saisonnier n'était pas constant dans le temps), alors (1.31) ne serait pas non plus une description raisonnable. La saisonnalité est effectivement un problème pratique difficile à résoudre dans les travaux appliqués avec des données chronologiques. Consulter le Chapitre 19.

Dans notre ouvrage, notre principal usage du Théorème FWL sera de faciliter la dérivation de résultats théoriques. Il est généralement plus aisé de traiter une équation comme (1.19), dans laquelle il y a une matrice unique de régresseurs, plutôt qu'une équation comme (1.18), où la matrice de régresseurs est partitionnée. Un exemple de la façon dont on peut utiliser le Théorème FWL pour dériver des résultats théoriques sera donné dans la Section 1.6.

## 1.5 CALCULER LES ESTIMATIONS OLS

Dans cette section nous discuterons brièvement de la façon dont les estimations OLS sont effectivement calculées par des ordinateurs. C'est un sujet avec lequel la plupart des étudiants en économétrie, et un certain nombre d'éconômètres, ne sont pas encore familiers. Dans la grande majorité des cas,

les programmes de régression bien formulés entraîneront des résultats fiables, et les économètres expérimentés ne se préoccupent donc pas de la manière dont on a obtenu ces résultats. Mais les programmes pour régression par OLS ne sont pas forcément tous bien spécifiés, et même les programmes les plus pointus peuvent rencontrer des difficultés si les données sont insuffisamment traitées. Nous croyons donc que les utilisateurs de logiciels de régressions par moindres carrés devraient être informés de ce que calcule exactement le programme. De plus, la méthode particulière pour la régression OLS sur laquelle nous porterons notre attention est intéressante d'un point de vue purement théorique.

Avant d'entamer une discussion sur la régression par moindres carrés, il nous faut expliquer comment les ordinateurs représentent les nombres réels, et comment cela influence la précision du calcul effectué sur de tels ordinateurs. A de rares exceptions près, les quantités pertinentes dans les régressions —  $\mathbf{y}$ ,  $\mathbf{X}$ ,  $\hat{\boldsymbol{\beta}}$ , et d'autres — sont des nombres réels plus souvent que des nombres entiers ou rationnels. En général, il faut une infinité de chiffres pour représenter un nombre réel avec exactitude, et cela est clairement impossible. Essayer de représenter chaque nombre, par autant de chiffres qu'il est nécessaire pour l'approximer avec "assez" de précision, équivaudrait à utiliser des quantités différentes de chiffres pour représenter des nombres différents; cela serait difficile à réaliser et ralentirait sensiblement les calculs. Alors les ordinateurs traitent généralement les nombres réels en les approximant par un nombre *fixe* de chiffres (ou, avec plus de précision, à l'aide de bits, qui correspondent aux chiffres en base deux). Mais pour manipuler des nombres qui peuvent se révéler très grands ou très petits, l'ordinateur les représente sous la forme de **nombres à virgule flottante**.<sup>6</sup>

L'idée fondamentale du nombre à virgule flottante est que n'importe quel nombre réel  $x$  peut aussi être écrit sous la forme

$$(b^c)m,$$

où  $m$ , la **mantisse** (ou partie décimale), est un nombre avec un signe compris entre zéro et un en valeur absolue,  $b$  est la **base** du système des nombres à virgule flottante, et  $c$  est l'**exposant**, qui est positif ou négatif. Ainsi, 663.725 peut s'écrire en base dix

$$0.663725 \times 10^3.$$

Conserver séparément la mantisse 663725 et l'exposant 3 offre un moyen pratique à l'ordinateur de conserver la valeur 663.725. L'avantage de ce système est que l'on peut conserver aussi facilement les très grands nombres et les très petits nombres, que les nombres de moindre amplitude; des valeurs telles que  $-0.192382 \times 10^{-23}$  et  $0.983443 \times 10^{17}$  peuvent être manipulées avec autant d'aisance que des nombres comme 3.42 ( $= 0.342 \times 10^1$ ).

<sup>6</sup> Notre introduction sur ce thème est nécessairement superficielle. Pour de plus amples détails, consulter Knuth (1981) ou Sterbenz (1974).



En pratique, les ordinateurs modernes n'utilisent pas la base dix; au lieu de cela, ils font un usage systématique des puissances de deux en tant que base (2, 4, 8 et 16 sont des valeurs employées en tant que bases pour les ordinateurs les plus largement utilisés), mais le principe demeure. La seule complication est que les nombres qui peuvent être représentés avec exactitude en base dix ne peuvent pas toujours (en fait, presque jamais) être représentés avec exactitude dans la base utilisée par l'ordinateur. Ainsi, il est très fréquent d'entrer une valeur, disons 6.8, dans un programme informatique, et de voir afficher 6.799999. Il s'est tout simplement passé que 6.8 a été converti dans la base que l'ordinateur utilise, puis converti à nouveau pour être affiché, et une erreur s'est introduite lors du déroulement du processus.

Seul un nombre restreint de nombres réels peut être représenté avec précision dans n'importe quelle base, et un plus petit nombre encore peut être représenté avec exactitude dans une base particulière. Alors la plupart des nombres ne peuvent être conservés que comme approximations. La précision et la finesse de l'approximation dépendent principalement du nombre de bits nécessaire pour conserver la mantisse. Typiquement, les programmeurs ont deux options à leur disposition: **la précision simple** et la **double précision**. Sur la plupart des ordinateurs, la mantisse d'un nombre à virgule flottante à précision simple pourra comporter au moins 21 ou 22 bits significatifs, alors que celle à double précision en comportera entre 50 et 54. Ceux-ci se transforment en 6 ou 7, et grossièrement en 15 ou 16 chiffres décimaux respectivement.<sup>7</sup>

Le problème majeur avec l'arithmétique en virgule flottante n'est pas que les nombres soient conservés comme approximations; après tout, la précision de la plupart des données économiques ne va pas au-delà de six chiffres de toute façon, de sorte que la simple précision est généralement suffisante pour représenter de telles données. La difficulté réelle est que lorsque l'on exécute des opérations arithmétiques sur des nombres en virgule flottante, des erreurs s'accumulent au cours des calculs. Comme nous aurons l'occasion de le constater plus loin, ces erreurs peuvent s'avérer tellement importantes que la réponse du calcul peut ne comporter aucun chiffre exact! Le type de problème le plus délicat survient lorsque des nombres de tailles et de signes différents

<sup>7</sup> Des ordinateurs différents conservent les nombres à virgule flottante de différentes manières. La plupart des ordinateurs modernes utilisent la précision simple avec 32 bits et la double précision avec 64 bits. Certains bits mémorisent l'exposant et le signe de la mantisse. Selon la base employée, des bits utilisés pour mémoriser la mantisse peuvent ne pas contenir toute l'information nécessaire, et de là, les nombres dans le texte pour les bits significatifs dans la mantisse. Un petit nombre d'ordinateurs utilise plus que 32 bits pour représenter les nombres en précision simple: 36, 48, 60, et 64 sont en usage. Sur ces ordinateurs à la fois l'arithmétique en précision simple et l'arithmétique à double précision seront plus précises que sur les ordinateurs 32 bit.

sont additionnés. Par exemple, considérons l'expression

$$2,393,121 - 1.0235 - 2,393,120, \quad (1.32)$$

qui donne comme résultat  $-0.0235$ . Supposons que l'on veuille tenter d'évaluer cette expression à l'aide d'un ordinateur qui travaille en base 10, et qui garde en mémoire six chiffres pour la mantisse. Si nous l'évaluons dans l'ordre dans lequel nous l'avons écrite, on obtient

$$0.239312 \times 10^7 - 0.102350 \times 10^1 - 0.239312 \times 10^7 \cong 0.000000 \times 10^1,$$

ou encore zéro, puisque,  $0.239312 \times 10^7 - 0.102350 \times 10^1 \cong 0.239312 \times 10^7$ , où " $\cong$ " correspond à l'égalité arithmétique au sens que lui donne l'ordinateur. De façon alternative, on pourrait changer l'ordre d'évaluation pour obtenir

$$0.239312 \times 10^7 - 0.239312 \times 10^7 - 0.102350 \times 10^1 \cong -0.102350 \times 10^1,$$

ou encore  $-1.0235$ . Mais aucun de ces résultats n'est acceptable pour la grande majorité des usages que l'on en fait.

De manière évidente, on peut rendre ce problème moins préoccupant en utilisant la double précision au lieu de la précision simple. Dans ce cas de figure, si nous avons fait usage des nombres à virgule flottante avec au moins onze chiffres pour la mantisse, nous aurions obtenu la réponse correcte. Mais il est clair que quelle que soit la quantité de chiffres utilisés dans nos opérations, des problèmes similaires à (1.32) surviendront constamment même si les nombres concernés sont plus petits ou plus grands, dès lors qu'il est impossible pour l'ordinateur de formuler une réponse acceptable. On fait référence à de telles difficultés en parlant de cas **insuffisamment conditionnés**, sans plus de précision.

La contrainte fondamentale dans l'arithmétique avec virgule flottante dont nous venons de discuter est d'une importance pratique considérable pour les économètres. Supposons que l'on veuille, par exemple, calculer la moyenne et la variance d'une série de nombres  $y_t$ ,  $t = 1, \dots, n$  issus d'un échantillon. Tout étudiant ayant fait de la statistique sait que

$$\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$$

et que l'estimation non biaisée de la variance des  $y_t$  est

$$\frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y})^2 \quad (1.33)$$

$$= \frac{1}{n-1} \left( \sum_{t=1}^n y_t^2 - n\bar{y}^2 \right). \quad (1.34)$$

L'égalité est vraie algébriquement, mais elle ne l'est plus lorsque l'on entame les calculs arithmétiques en virgule flottante. La première régression, (1.33), peut être évaluée avec une précision raisonnable, tant que l'amplitude entre les  $y_t$  et  $\bar{y}$  n'est pas très importante. L'expression (1.34), toutefois, nécessite de soustraire  $n\bar{y}^2$  à  $\sum y_t^2$ , et lorsque  $\bar{y}$  est grande par rapport à la variance des  $y_t$ , la différence entre les deux grandeurs peut être très importante. Ainsi dans cette situation, l'expression (1.34) peut produire un calcul de la variance très approximatif. De telles expressions sont souvent considérées comme **numériquement instables**, parce qu'elles sont prédisposées à certaines erreurs lorsqu'elles sont évaluées à l'aide de l'arithmétique à virgule flottante.

L'ampleur des difficultés numériques que l'on rencontre peut être aperçue et illustrée par un exemple numérique simple. On génère tout d'abord 1000 nombres pseudo-aléatoires d'après une distribution normale (consulter le Chapitre 21), et nous les normalisons, de manière à ce que la moyenne d'échantillonnage soit exactement  $\mu$ , et la variance d'échantillonnage soit exactement égale à l'unité.<sup>8</sup> Puis l'on calcule la variance de l'échantillon pour quelques valeurs de  $\mu$ , en utilisant la précision simple et la double précision pour les expressions (1.33) et (1.34). Les résultats, exprimés comme la valeur absolue de la *différence* entre la variance calculée et la véritable variance égale à l'unité (et présentée avec la notation à virgule flottante puisque ces différences peuvent avoir des amplitudes très variables) sont collectés dans le Tableau 1.1.<sup>9</sup>

Cet exemple illustre deux phénomènes importants. Le premier d'entre eux, exception faite lorsque  $\mu = 0$ , de sorte qu'aucun problème numérique n'apparaissait dans l'une et l'autre des formules, nous confirme ce que notre discussion suggérait, c'est-à-dire que l'expression (1.33) offre des résultats plus fins que l'expression (1.34). Le second montre que l'arithmétique à double précision entraîne des résultats plus précis que l'arithmétique à précision simple. Il y a un avantage à évaluer l'expression numériquement instable (1.34) en double précision plutôt que l'expression numériquement stable (1.33) en

<sup>8</sup> En réalité, la normalisation n'était pas *exacte*, mais elle était extrêmement précise par l'usage de la **quadruple précision**, qui est environ deux fois plus fine que la double précision. L'arithmétique en, quadruple précision n'est pas disponible sur de nombreux ordinateurs (en particulier sur les petits modèles) et elle est particulièrement plus lente que l'arithmétique à double précision, mais elle peut fournir des résultats encore plus fins. Nous sommes persuadés que les séries avec lesquelles nous avons débuté ont vraiment une moyenne  $\mu$  et une variance unitaire à au moins 30 chiffres décimaux.

<sup>9</sup> Tous les calculs furent programmés en FORTRAN VS et effectués sur un ordinateur IBM travaillant sous VM/CMS. Sur d'autres matériels, les résultats seraient légèrement différents, même si les nombres flottants à simple ou double précision sont représentés respectivement en 32 ou 64 bits. En particulier, la plupart des ordinateurs personnels fourniraient des résultats plus précis.

**Tableau 1.1** Erreurs Absolues du Calcul de la Variance de l'Échantillon

$\mu$	(1.33) simple	(1.33) double	(1.34) simple	(1.34) double
0	$0.880 \times 10^{-4}$	$0.209 \times 10^{-13}$	$0.880 \times 10^{-4}$	$0.209 \times 10^{-13}$
10	$0.868 \times 10^{-4}$	$0.207 \times 10^{-13}$	$0.126 \times 10^0$	$0.281 \times 10^{-11}$
$10^2$	$0.553 \times 10^{-4}$	$0.208 \times 10^{-13}$	$0.197 \times 10^1$	$0.478 \times 10^{-9}$
$10^3$	$0.756 \times 10^{-3}$	$0.194 \times 10^{-13}$	$0.410 \times 10^2$	$0.859 \times 10^{-8}$
$10^4$	$0.204 \times 10^0$	$0.179 \times 10^{-13}$	$0.302 \times 10^4$	$0.687 \times 10^{-6}$
$10^5$	$0.452 \times 10^2$	$0.180 \times 10^{-14}$	$0.733 \times 10^6$	$0.201 \times 10^{-3}$

précision simple. Mais les meilleurs résultats sont obtenus bien évidemment en évaluant l'expression (1.33) en faisant usage de calculs arithmétiques en double précision. En adoptant cette approche, la précision est excellente pour tout l'éventail de valeurs de  $\mu$  proposé dans le Tableau (elle se détériore progressivement lorsque  $\mu$  dépasse  $10^6$  de manière significative). Par contraste, les deux formules mènent à un non-sens lorsqu'elle sont calculées pour  $\mu$  égal à  $10^5$ , en précision simple, et (1.34) donne un résultat très approximatif même lorsque l'on utilise la double précision.

Nous espérons que cet exemple mettra en lumière le fait que toute tentative de calcul des estimations pour l'évaluation d'expressions algébriques standards, sans avoir conscience des conditions relatives à la précision de la machine et à la stabilité numérique des expressions, est en fait très imprudente. Le meilleur moyen qu'il convient d'adopter est de toujours faire usage de logiciels écrits par des experts qui ont intégré de telles considérations. Si de tels logiciels ne sont pas disponibles, alors l'utilisateur occasionnel devrait utiliser la double précision, avec une précision accrue (si possible) pour des opérations plus délicates. Comme l'exemple l'indique, même des procédures numériquement stables peuvent fournir des non-sens avec une précision simple en 32 bits si les données sont insuffisamment conditionnées.<sup>10</sup>

Maintenant, retournons au sujet principal de cette section, qui reste le calcul des estimations par moindres carrés ordinaires. De nombreuses références solides sont disponibles sur ce sujet—consulter, entre autres, Chambers (1977), Kennedy et Gentle (1980), Maindonald (1984), Farebrother (1988), et Golub et Van Loan (1989)—et nous nous permettrons donc de ne pas entrer dans trop de détails.

<sup>10</sup> Un exemple classique est celui du calcul d'un produit intérieur. Celui-ci est généralement réalisé avec une boucle d'itérations, et la valeur du produit intérieur est mémorisée dans une variable scalaire. Les propriétés numériques d'une telle procédure peuvent être largement améliorées en conservant cette variable dans la précision la plus grande possible, même si les calculs ultérieurs sont réalisés avec une précision moindre.

Le moyen évident d'obtenir  $\hat{\beta}$  est premièrement de construire une matrice des sommes des carrés et des produits croisés des régresseurs et de la régressande, ou de manière équivalente la matrice  $\mathbf{X}^\top \mathbf{X}$  et le vecteur  $\mathbf{X}^\top \mathbf{y}$ . Il faudrait ensuite inverser la première par une technique générale d'inversion matricielle, et postmultiplier  $(\mathbf{X}^\top \mathbf{X})^{-1}$  par  $\mathbf{X}^\top \mathbf{y}$ . Malheureusement, cette procédure recouvre tous les défauts et les inconvénients de l'expression (1.34). Cela pourrait aboutir avec satisfaction si la double précision était utilisée tout au long de la procédure, toutes les colonnes de  $\mathbf{X}$  étant d'amplitude similaire, et la matrice  $\mathbf{X}^\top \mathbf{X}$  n'étant pas singulière, mais cela serait peu recommandé dans le cas général.

Il y a deux approches majeures pour calculer des estimations par moindres carrés que l'on peut conseiller. L'une est pour l'essentiel une version plus sophistiquée de celle que nous venons de décrire. Elle nécessite également la construction d'une matrice des sommes des carrés et des produits croisés des régresseurs et de la régressande, mais d'une manière telle que les problèmes numériques soient réduits. Un moyen efficace d'éviter de tels problèmes est de soustraire la moyenne des variables avant d'effectuer la mise au carré et les produits croisés. Pratiquer de cette façon nécessite de faire deux passages par les données, ce qui peut malgré tout se révéler désagréable si l'ensemble des données est trop étendu pour être assimilé par la mémoire vive de l'ordinateur, et une technique alternative au moins aussi précise peut être choisie; consulter Maindonald (1984). Avec l'une ou l'autre des deux techniques, la matrice originelle des sommes des carrés et des produits croisés est reconstituée, et les équations normales sont résolues par la suite à l'aide, soit d'une variante de la décomposition de Cholesky, soit d'une élimination Gaussienne conventionnelle. Il est important pour la précision numérique de résoudre les équations normales, qui peuvent produire l'inverse de  $\mathbf{X}^\top \mathbf{X}$  comme un produit dérivé, plutôt que d'inverser en premier  $\mathbf{X}^\top \mathbf{X}$  et de construire ensuite  $\hat{\beta}$  par multiplication. L'utilisation de la décomposition de Cholesky nous permet de tirer avantage du fait que  $\mathbf{X}^\top \mathbf{X}$  est une matrice symétrique définie positive, et peut donc se révéler d'une certaine manière plus performante que l'élimination Gaussienne. Pour plus de détails, consulter Maindonald (1984).

L'autre approche pour calculer les estimations par moindres carrés impliquent la recherche d'une **base orthonormale** pour le sous-espace engendré par les colonnes de  $\mathbf{X}$ . Il s'agit encore d'une matrice de dimension  $n \times k$ , disons  $\mathbf{Q}$ , avec les propriétés d'égalité entre  $\mathcal{S}(\mathbf{Q}) = \mathcal{S}(\mathbf{X})$ , et  $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ . Cette approche est celle sur laquelle nous porterons notre attention, d'une part parce qu'elle fournit les résultats les plus précis (bien qu'entraînant un coût non négligeable en terme de temps d'exécution pour la machine), et d'autre part parce qu'elle s'avère intéressante du point de vue théorique. C'est la seule approche que conseille Chambers (1977), et c'est également la seule que recommande Maindonald (1984) lorsque la précision est le critère de la plus haute importance. Les lecteurs devraient se reporter à ces références pour tous les détails que nous ne traiterons pas.

Pour toute matrice de régresseurs  $\mathbf{X}$  de rang  $k$ , il est possible de trouver une matrice  $\mathbf{Q}$  de dimension  $n \times k$  et une matrice triangulaire supérieure  $\mathbf{R}$  de dimension  $k \times k$ , pour lesquelles on a

$$\mathbf{X} = \mathbf{Q}\mathbf{R} \quad \text{et} \quad \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}. \quad (1.35)$$

La seconde condition implique que les colonnes de  $\mathbf{Q}$  soient orthonormales: elles ont toutes une longueur euclidienne égale à l'unité, et sont mutuellement orthogonales. Le fait que  $\mathbf{R}$  soit triangulaire implique que les colonnes de  $\mathbf{Q}$  sont reliées de manière récursive. La première colonne de  $\mathbf{Q}$  est tout simplement la première colonne  $\mathbf{X}$ , avec une nouvelle échelle pour avoir une longueur unitaire; la deuxième colonne de  $\mathbf{Q}$  est une transformation linéaire des deux premières colonnes de  $\mathbf{X}$  qui est orthogonale à la première colonne de  $\mathbf{Q}$ , et dont la longueur est encore égale à l'unité; et ainsi de suite. Il y a plusieurs manières de trouver  $\mathbf{Q}$  et  $\mathbf{R}$ , dont les deux les plus importantes sont la méthode de Gram-Schmidt et la transformation de Householder. Ce sont des techniques qui sont assez semblables dans leurs calculs, et l'on pourra en trouver une description dans les références de Chambers et Maindonald que nous avons données. Les deux techniques sont simples et donc, ne nécessitent qu'une écriture en machine assez compacte, pourvu que la méthode disponible puisse traiter les cas dans lesquels  $\mathbf{X}$  n'est pas de plein rang (ou apparaît à l'ordinateur n'être pas de plein rang).

Décider si  $\mathbf{X}$  est, ou n'est pas, de plein rang est un problème délicat pour tout algorithme basé sur les moindres carrés, puisqu'à cause des erreurs d'arrondi, les ordinateurs ne peuvent pas détecter la différence de manière fiable entre des nombres qui sont véritablement zéro, et des nombres qui en sont très proches. C'est une des raisons pour lesquelles il est important que les données soient exprimées dans la même échelle. Lorsque  $m$  colonnes de  $\mathbf{X}$  sont linéairement dépendantes des autres colonnes de  $\mathbf{X}$ , il est nécessaire de modifier l'algorithme de manière à ce que  $\mathbf{Q}$  possède  $k - m$  colonnes et  $\mathbf{R}$  soit de dimension  $(k - m) \times k$ . Les estimations  $\hat{\beta}$  sont ensuite calculées de manière unique en initialisant arbitrairement à zéro les coefficients des  $m$  régresseurs linéairement dépendantes.

Supposons que l'on ait trouvé  $\mathbf{Q}$  et  $\mathbf{R}$  telles qu'elles satisfont (1.35). Il est alors très facile de calculer les quantités qui nous intéressent. La fonction de régression  $\mathbf{X}\beta$  peut alors s'écrire  $\mathbf{Q}\mathbf{R}\beta = \mathbf{Q}\gamma$ , et il est aisé de constater que l'estimation OLS de  $\gamma$  est

$$\hat{\gamma} = (\mathbf{Q}^\top \mathbf{Q})^{-1} \mathbf{Q}^\top \mathbf{y} = \mathbf{Q}^\top \mathbf{y},$$

qu'il est extrêmement simple d'évaluer. Il est également aussi aisé d'estimer les valeurs ajustées  $\mathbf{Q}\hat{\gamma}$  et les résidus

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{Q}\hat{\gamma} = \mathbf{y} - \mathbf{Q}\mathbf{Q}^\top \mathbf{y}. \quad (1.36)$$

Ainsi, si seuls les résidus et/ou les valeurs ajustées nous intéressent, il n'est pas du tout nécessaire de calculer  $\hat{\beta}$ .

Notons qu'à partir de (1.36), les matrices de projection  $P_X$  et  $M_X$  sont égales respectivement à  $QQ^\top$  et  $I - QQ^\top$ . La simplicité de ces expressions découle du fait que  $Q$  représente une base orthonormale de  $\mathcal{S}(X)$ . Géométriquement, rien n'aurait été modifié si nous avions utilisé  $Q$  au lieu de  $X$  comme matrice de régresseurs dans chacune des figures que nous avons dessinées, puisque  $\mathcal{S}(Q) = \mathcal{S}(X)$ . Si nous devons montrer les colonnes de  $Q$  sur chaque figure, chacune serait un point dans  $\mathcal{S}(X)$  localisé sur la sphère unitaire (c'est-à-dire sur la sphère centrée sur l'origine, et de rayon égal à l'unité) et formerait un angle droit avec les points représentant les autres colonnes de  $Q$ .

Pour rendre possible le calcul de  $\hat{\beta}$  et de  $(X^\top X)^{-1}$ , qui seront souvent les principales grandeurs qui nous intéresseront, nous faisons usage du fait que  $\hat{\beta} = R^{-1}\hat{\gamma}$  et

$$(X^\top X)^{-1} = (R^\top Q^\top QR)^{-1} = (R^\top R)^{-1} = R^{-1}(R^{-1})^\top.$$

Ainsi, dès que l'on a obtenu  $R^{-1}$ , il est très facile de calculer les estimations moindres carrés  $\hat{\beta}$ , et leur matrice de covariance estimée (consulter Chapitre 2). Puisque  $R$  est une matrice triangulaire, son inverse se calcule facilement et sans coût important pour l'ordinateur: nous n'avons même pas besoin de vérifier qu'elle n'est pas singulière, puisque  $R$  ne sera pas de plein rang que si  $X$  n'est pas de plein rang, ce qui se découvrira et se traitera lors de la construction de  $Q$  et de  $R$ .

La partie la plus coûteuse de ces procédures pour l'ordinateur sera le plus souvent la construction des matrices  $Q$  et  $R$  à partir de  $X$ . Elle nécessite un nombre d'opérations arithmétiques qui est approximativement proportionnel à  $nk^2$ . La construction des matrices des sommes des carrés et des produits croisés, qui est le point de départ de la méthode ayant pour objet la résolution des équations normales, nécessite également un nombre d'opérations proportionnel à  $nk^2$ , bien que le facteur de proportionnalité soit plus restreint. Ainsi la régression linéaire peut s'avérer très coûteuse lorsque le nombre de régresseurs et/ou la taille de l'échantillon sont très élevés, quelle que soit la méthode choisie. Si l'on est amené à estimer plusieurs régressions sur la base du même ensemble de données, il est intéressant de réduire le coût en n'effectuant les calculs onéreux qu'une seule fois. Plusieurs logiciels de régression permettent à l'utilisateur de construire préalablement les matrices des sommes des carrés et des produits croisés pour toutes les variables de l'ensemble de données, et de calculer ensuite des estimations de diverses régressions, en extrayant les lignes et les colonnes pertinentes pour la résolution des équations normales. Si c'est cette approche qui est envisagée, alors il devient particulièrement important d'échelonner les données, de manière à ce que les divers régresseurs ne soient pas trop divergents en moyenne et en variance.

## 1.6 OBSERVATIONS INFLUENTES ET L'EFFET LEVIER

Chaque élément du vecteur des estimations OLS de  $\hat{\beta}$  est tout simplement une moyenne *pondérée* des éléments du vecteur  $\mathbf{y}$ . Pour l'apercevoir, définissons  $\mathbf{c}_i$  comme la  $i^{\text{ième}}$  ligne de la matrice  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  et observons à partir de (1.04) que

$$\hat{\beta}_i = \mathbf{c}_i \mathbf{y}.$$

Puisque chaque élément de  $\hat{\beta}$  est une moyenne pondérée, certaines observations auront plus de poids, plus d'influence sur  $\hat{\beta}$  que les autres. Si une, ou un petit nombre d'observations sont extrêmement **influentes**, dans le sens où en les éliminant, cela modifierait très fortement les éléments de  $\hat{\beta}$ , l'économetre prudent cherchera normalement à examiner consciencieusement les données. Il est fort possible que ces observations influentes soient erronées, ou pour telle ou telle raison, atypiques, par rapport aux autres observations de l'échantillon. Comme nous le verrons, même une observation erronée isolée peut avoir un impact considérable sur  $\hat{\beta}$  dans certains cas, de sorte qu'il devient extrêmement important de déceler et de corriger de telles observations si elles sont effectivement influentes. Même si les données sont correctes, l'interprétation des résultats peut être modifiée substantiellement s'il est prouvé qu'une ou un petit nombre d'observations sont responsables au premier chef de ces résultats, et tout particulièrement si ces observations divergent systématiquement du reste des données.

La littérature traitant de la détection des observations influentes est assez récente, et elle n'a pas encore été entièrement assimilée en économétrie et pour l'ensemble des logiciels disponibles. Les références à consulter sur ce sujet sont Belsley, Kuh, et Welsch (1980), Cook et Weisberg (1982), et Krasker, Kuh, et Welsch (1983). Dans cette section, nous introduirons seulement quelques concepts fondamentaux, et les résultats avec lesquels tout économètre devrait se sentir familier. La démonstration de ces résultats offre un bon exemple de l'utilité du Théorème FWL.

L'effet d'une observation isolée sur  $\hat{\beta}$  peut être apprécié en comparant  $\hat{\beta}$  avec  $\hat{\beta}^{(t)}$ , l'estimation de  $\beta$  obtenue par OLS sur un échantillon dont on a ôté l'observation  $t$ . La différence entre  $\hat{\beta}$  et  $\hat{\beta}^{(t)}$  s'avèrera dépendre de façon critique de la quantité

$$h_t \equiv \mathbf{X}_t (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top, \quad (1.37)$$

qui est le  $t^{\text{ième}}$  élément diagonal de la matrice  $\mathbf{P}_X$ . La notation  $h_t$  dérive du fait que l'on se réfère quelquefois à  $\mathbf{P}_X$  en tant que **matrice chapeau**; parce que  $\hat{\mathbf{y}} \equiv \mathbf{P}_X \mathbf{y}$ ,  $\mathbf{P}_X$  “dépose un chapeau sur”  $\mathbf{y}$ . Notons que  $h_t$  dépend uniquement de la matrice  $\mathbf{X}$  des régresseurs et non pas de la régressande  $\mathbf{y}$ .

La situation s'éclaire lorsque l'on reformule  $h_t$  comme

$$h_t = \mathbf{e}_t^\top \mathbf{P}_X \mathbf{e}_t = \|\mathbf{P}_X \mathbf{e}_t\|^2, \quad (1.38)$$



où  $\mathbf{e}_t$  représente le vecteur à  $n$  composantes avec un 1 à la  $t^{\text{ième}}$  position et des 0 partout ailleurs. L'expression (1.38) découle de (1.37), de la définition de  $\mathbf{P}_X$ , et du fait que  $\mathbf{e}_t^\top \mathbf{X} = \mathbf{X}_t$ . L'expression la plus à droite montre que  $h_t$  est la norme au carré d'un certain vecteur, ce qui garantit que  $h_t \geq 0$ . De plus, comme  $\|\mathbf{e}_t\| = 1$ , et puisque la longueur du vecteur  $\mathbf{P}_X \mathbf{e}_t$  ne peut dépasser celle de  $\mathbf{e}_t$ , alors il doit être vérifié que  $h_t = \|\mathbf{P}_X \mathbf{e}_t\|^2 \leq 1$ . Ainsi (1.38) assure que

$$0 \leq h_t \leq 1. \quad (1.39)$$

Supposons que  $\hat{u}_t$  représente le  $t^{\text{ième}}$  élément du vecteur des résidus des moindres carrés  $\mathbf{M}_X \mathbf{y}$ . On peut alors établir le résultat fondamental selon lequel

$$\hat{\boldsymbol{\beta}}^{(t)} = \hat{\boldsymbol{\beta}} - \left( \frac{1}{1 - h_t} \right) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top \hat{u}_t. \quad (1.40)$$

Il est clair d'après cette expression que lorsque  $\hat{u}_t$  est grand et/ou lorsque  $1 - h_t$  est petit, l'effet de la  $t^{\text{ième}}$  observation sur au moins quelques éléments de  $\hat{\boldsymbol{\beta}}$  est vraisemblablement substantiel. Nous démontrerons ce résultat plus tard.

Il peut se révéler très utile de constater combien l'omission de l'observation  $t$  de la régression influe sur les valeurs ajustées de cette observation. De (1.40), il découle directement que

$$\mathbf{X}_t \hat{\boldsymbol{\beta}}^{(t)} = \mathbf{X}_t \hat{\boldsymbol{\beta}} - \left( \frac{h_t}{1 - h_t} \right) \hat{u}_t. \quad (1.41)$$

Dans la pratique, l'utilisation de (1.40) peut se révéler fastidieuse pour vérifier si chaque observation est ou n'est pas influente en regardant si son omission a un impact non négligeable sur l'un des éléments de  $\hat{\boldsymbol{\beta}}$ . Mais une observation est certainement très influente si son omission affecte sa propre valeur ajustée. De (1.41), nous constatons que la modification de la  $t^{\text{ième}}$  valeur ajustée consécutive à l'omission de l'observation  $t$  est  $-\hat{u}_t h_t / (1 - h_t)$ . Il découle immédiatement que la modification du  $t^{\text{ième}}$  résidu est

$$\left( \frac{h_t}{1 - h_t} \right) \hat{u}_t. \quad (1.42)$$

Un moyen simple de détecter les observations qui sont influentes, dans le sens où elles affectent les valeurs ajustées et les résidus, est donc de tracer l'expression (1.42) en fonction de  $t$ .

Ces résultats suggèrent un examen plus approfondi des grandeurs  $h_t$ . Nous avons déjà établi dans (1.39) que les  $h_t$  sont tous compris entre zéro et un. En fait, leur somme atteint  $k$ , un résultat que l'on montre aisément en faisant usage des propriétés de la trace d'une matrice (consulter Annexe A):

$$\begin{aligned} \sum_{t=1}^n h_t &= \text{Tr}(\mathbf{P}_X) = \text{Tr}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \\ &= \text{Tr}((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}) = \text{Tr}(\mathbf{I}_k) = k. \end{aligned}$$

Ainsi, en moyenne, un  $h_t$  égale  $k/n$ . Lorsqu'il existe un terme constant, chaque  $h_t$  est au plus égal à  $1/n$ , une propriété que l'on décèle facilement à partir de (1.38), puisque si  $\mathbf{X}$  n'était constituée que d'un vecteur dont les composantes seraient des constantes,  $\mathbf{e}_t^\top \mathbf{P}_X \mathbf{e}_t$  serait égal à  $1/n$ . Même s'il n'y a pas de terme constant,  $h_t$  n'est jamais nul, à moins que  $\mathbf{X}_t$  ne soit composée que de zéros. Cependant, il est évidemment possible de rencontrer des  $h_t$  égaux à l'unité. Supposons par exemple que l'une des colonnes de  $\mathbf{X}$  soit la variable muette représentée par  $\mathbf{e}_t$ . Alors,  $h_t = \mathbf{e}_t^\top \mathbf{P}_X \mathbf{e}_t = \mathbf{e}_t^\top \mathbf{e}_t = 1$ .

Il est intéressant de constater ce qu'il advient si l'on ajoute une variable muette représentée par  $\mathbf{e}_t$  à une régression. Il apparaît que  $\hat{u}_t$  est nul et que l'observation  $t$  n'aura aucun effet sur les coefficients, sauf celui qui correspond à la variable muette. Celui-ci prend tout simplement la valeur nécessaire pour établir que  $\hat{u}_t = 0$ , et les coefficients restant sont ceux qui minimisent la fonction SSR pour les  $n - 1$  observations restantes. Ces résultats découlent aisément de l'usage du Théorème FWL.

Considérons les deux régressions suivantes où pour alléger les notations, les données ont été ordonnées de manière à ce que l'observation  $t$  soit la dernière, et  $\mathbf{y}_{(t)}$  et  $\mathbf{X}_{(t)}$  désignent les  $n - 1$  premières lignes de  $\mathbf{y}$  et de  $\mathbf{X}$ , respectivement:

$$\begin{bmatrix} \mathbf{y}_{(t)} \\ y_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(t)} \\ \mathbf{X}_t \end{bmatrix} \boldsymbol{\beta} + \text{résidus}, \quad (1.43)$$

et

$$\begin{bmatrix} \mathbf{y}_{(t)} \\ y_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(t)} & \mathbf{0} \\ \mathbf{X}_t & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \alpha \end{bmatrix} + \text{résidus}. \quad (1.44)$$

La régression (1.43) est tout simplement la régression de  $\mathbf{y}$  sur  $\mathbf{X}$ , qui permet d'élaborer les estimations des paramètres  $\hat{\boldsymbol{\beta}}$  et les résidus des moindres carrés. La régression (1.44) est la régression (1.43) à laquelle nous avons ajouté le régresseur  $\mathbf{e}_t$ . D'après le Théorème FWL, les estimations de  $\boldsymbol{\beta}$  données par (1.44) doivent être équivalentes à celles données par la régression

$$\mathbf{M}_t \begin{bmatrix} \mathbf{y}_{(t)} \\ y_t \end{bmatrix} = \mathbf{M}_t \begin{bmatrix} \mathbf{X}_{(t)} \\ \mathbf{X}_t \end{bmatrix} \boldsymbol{\beta} + \text{résidus}, \quad (1.45)$$

où  $\mathbf{M}_t$  est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{e}_t)$ . Multiplier n'importe quel vecteur par  $\mathbf{M}_t$  annule purement et simplement la dernière composante de ce vecteur. Ainsi la régression (1.45) est aussi

$$\begin{bmatrix} \mathbf{y}_{(t)} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{(t)} \\ \mathbf{0} \end{bmatrix} \boldsymbol{\beta} + \text{résidus}. \quad (1.46)$$

La dernière observation pour laquelle la régressande et les régresseurs sont nuls, n'a à l'évidence aucun effet sur les estimations des paramètres, donc (1.46) est équivalente à la régression de  $\mathbf{y}_{(t)}$  sur  $\mathbf{X}_{(t)}$ , et doit par conséquent fournir les mêmes estimations OLS  $\hat{\boldsymbol{\beta}}^{(t)}$ . Pour la régression (1.46) le résidu

rattaché à l'observation  $t$  est évidemment nul; le Théorème FWL implique alors que le résidu de l'observation  $t$  de la régression (1.44) doit aussi être nul, ce qui établit que  $\hat{\alpha}$  doit être égal à  $y_t - \mathbf{X}_t \hat{\boldsymbol{\beta}}^{(t)}$ .

Ces résultats doivent rendre plus facile la démonstration de (1.40) et de (1.41) que nous avons établies plus tôt sans preuve. Les lecteurs qui ne sont pas friands de démonstration désireraient sans doute sauter les trois paragraphes qui vont suivre. Quoi qu'il en soit, ces démonstrations illustrent la puissance de l'algèbre combinée au Théorème FWL et à une compréhension de la géométrie impliquée. Il est donc intéressant et instructif de la comparer avec des preuves traditionnelles telles que celles fournies dans l'Annexe 2A de Belsley, Kuh, et Welsch (1980).

Par l'usage des résultats que nous venons de démontrer, nous voyons que les valeurs ajustées de (1.44) sont  $\mathbf{X} \hat{\boldsymbol{\beta}}^{(t)} + \mathbf{e}_t \hat{\alpha}$ , alors que celles de la régression (1.43) sont  $\mathbf{X} \hat{\boldsymbol{\beta}}$ . Moins leur différence, qui est égale à la différence entre les résidus de (1.43) et de (1.44), est  $\mathbf{X} \hat{\boldsymbol{\beta}} - \mathbf{X} \hat{\boldsymbol{\beta}}^{(t)} - \mathbf{e}_t \hat{\alpha}$ . En prémultipliant cette différence par  $\mathbf{M}_X$  (ou, bien sûr,  $\mathbf{X} \equiv [\mathbf{X}_{(t)} \vdots \mathbf{X}_t]$ ) il vient

$$\begin{aligned} \mathbf{M}_X (\mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(t)}) - \mathbf{e}_t \hat{\alpha}) &= \mathbf{M}_X (\hat{\mathbf{u}}^{(t)} - \hat{\mathbf{u}}) \\ &= \hat{\mathbf{u}}^{(t)} - \hat{\mathbf{u}} = -\mathbf{M}_X \mathbf{e}_t \hat{\alpha}, \end{aligned} \quad (1.47)$$

où  $\hat{\mathbf{u}}^{(t)}$  désigne le résidu de (1.44). On prémultiplie maintenant chaque membre de l'égalité de la seconde ligne de (1.47) par  $-\mathbf{e}_t^\top$ . Comme nous l'avons montré plus tôt,  $\hat{u}_t^{(t)} = 0$ , de sorte que  $\mathbf{e}_t^\top \hat{\mathbf{u}}^{(t)} = 0$ , et ainsi le résultat de cette prémultiplication est

$$\mathbf{e}_t^\top \mathbf{M}_X \mathbf{e}_t \hat{\alpha} = \hat{u}_t. \quad (1.48)$$

Par définition  $\mathbf{e}_t^\top \mathbf{M}_X \mathbf{e}_t = 1 - h_t$ , et donc (1.48) implique que

$$\hat{\alpha} = \frac{\hat{u}_t}{1 - h_t}. \quad (1.49)$$

Puisque  $\hat{\alpha}$  est  $y_t - \mathbf{X}_t \hat{\boldsymbol{\beta}}^{(t)}$ , le résultat (1.49) nous offre ce dont nous avons besoin. La modification dans le résidu  $t$  occasionnée par l'omission de l'observation  $t$  doit être

$$\hat{\alpha} - \hat{u}_t = \frac{\hat{u}_t}{1 - h_t} - \hat{u}_t = \left( \frac{h_t}{1 - h_t} \right) \hat{u}_t,$$

qui correspond à l'expression (1.42). Il faut soustraire la modification du résidu  $t$  de sorte que

$$\mathbf{X}_t \hat{\boldsymbol{\beta}}^{(t)} - \mathbf{X}_t \hat{\boldsymbol{\beta}} = - \left( \frac{h_t}{1 - h_t} \right) \hat{u}_t$$

d'où le résultat (1.41) découle immédiatement.

L'on peut maintenant dériver (1.40). En faisant usage de (1.47), on constate que

$$\hat{\mathbf{u}} - \hat{\mathbf{u}}^{(t)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)} + \mathbf{e}_t\hat{\alpha} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

En prémultipliant cette quantité par  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  on obtient

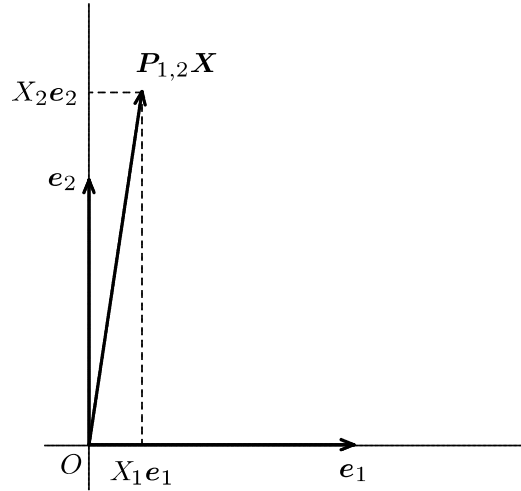
$$\begin{aligned} \mathbf{0} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}}^{(t)} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{e}_t \hat{\alpha} \\ &= \hat{\boldsymbol{\beta}}^{(t)} - \hat{\boldsymbol{\beta}} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{e}_t \hat{\alpha}, \end{aligned}$$

où le membre de gauche est nul parce que  $\hat{\mathbf{u}} - \hat{\mathbf{u}}^{(t)}$  se situe dans  $\mathcal{S}^\perp(\mathbf{X})$ . En résolvant, pour  $\hat{\boldsymbol{\beta}}^{(t)}$  par l'intermédiaire de (1.49) on obtient le résultat fondamental (1.40):

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(t)} &= \hat{\boldsymbol{\beta}} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{e}_t (1 - h_t)^{-1} \hat{u}_t \\ &= \hat{\boldsymbol{\beta}} - \left( \frac{1}{1 - h_t} \right) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top \hat{u}_t. \end{aligned}$$

Nous avons vu que les quantités  $h_t$  ne doivent pas dépasser l'unité pas plus qu'elles ne peuvent être négatives, et qu'en moyenne elles valent  $k/n$ . Nous avons vu également que l'omission de l'observation  $t$  avait d'autant plus d'impact sur  $\hat{\boldsymbol{\beta}}$  que  $h_t$  était relativement fort, à moins que  $\hat{u}_t$  ne soit proche de zéro. Ainsi les quantités  $h_t$  peuvent servir d'indicateur de **puissance**, ou d'effet potentiel sur  $\hat{\boldsymbol{\beta}}$ , pour chacune des observations. On dit des observations pour lesquelles  $h_t$  est relativement important (par exemple supérieur à  $2k/n$ ) qu'elles ont un poids important ou que ce sont des points à forte pondération. Si tous les  $h_t$  étaient égaux à  $k/n$ , comme cela serait le cas si le régresseur était une constante, alors chaque observation aurait le même poids. On désigne souvent cette situation par le terme de **modèle équilibré**, et c'est la situation la plus enviable, mais puisqu'en économétrie la construction de la matrice  $\mathbf{X}$  est rarement contrôlable par l'investigateur, c'est une situation peu commune. Notons que  $t$  peut avoir un poids important sans pour autant être influent si  $h_t$  est élevé mais  $\hat{u}_t$  petit. Un point à forte pondération possède une influence *potentielle*, mais la réalisation effective de cette influence dépend de  $y_t$ .

L'un des moyens de prendre conscience de la pondération est de faire usage du fait que  $h_t = \mathbf{e}_t^\top \mathbf{P}_X \mathbf{e}_t$ . De cette propriété nous voyons que  $h_t$  sera fort si la matrice de régresseurs  $\mathbf{X}$  a un poids explicatif important de la variable muette  $\mathbf{e}_t$  et  $h_t$  sera faible dans le cas contraire. Imaginons qu'il nous faille remplacer  $\mathbf{y}$  par  $\mathbf{y} + \delta \mathbf{e}_t$ , ajoutant  $\delta$  à  $y_t$  pour l'observation  $t$  uniquement. Si l'observation  $t$  avait un poids faible ( $h_t$  petit),  $\hat{\boldsymbol{\beta}}$  ne serait pas beaucoup modifié et  $\hat{u}_t$  devrait varier d'au moins la variation enregistrée par  $y_t$ , c'est-à-dire  $\delta$ . A contrario, si l'observation  $t$  avait un poids important ( $h_t$  élevé) au moins un élément de  $\hat{\boldsymbol{\beta}}$  varierait largement et  $\hat{u}_t$  varierait de beaucoup moins que  $\delta$ . Ainsi plus le poids de l'observation  $t$  sera important, plus la valeur de  $y_t$  aura un effet sur  $\hat{\boldsymbol{\beta}}$ , et moins elle aura d'impact sur  $\hat{u}_t$ .



**Figure 1.8** Le poids relativement plus important de l'observation 2

La Figure 1.8 illustre le cas où  $k = 1$ , c'est-à-dire celui où il n'y a qu'un seul régresseur et où la seconde observation a beaucoup plus d'importance que la première. Exceptionnellement les deux axes horizontaux et verticaux représentent effectivement les deux observations, et ainsi ce que nous observons, c'est la projection du vecteur représentant le régresseur  $\mathbf{X}$  sur l'espace engendré par les deux vecteurs  $\mathbf{e}_1$  et  $\mathbf{e}_2$  correspondant aux deux premières observations comme l'indique le fait que les deux vecteurs sont situés sur les deux axes. La projection de  $\mathbf{X}$  est le vecteur noté  $\mathbf{P}_{1,2}\mathbf{X}$  sur la figure. Alors qu'ici il n'y a qu'un régresseur, la quantité  $h_t$  devient simplement  $X_t^2 / \|\mathbf{X}\|^2$ , comme on peut le constater à partir de (1.38). Ainsi le ratio  $h_2$  sur  $h_1$ , ou le poids relatif de l'observation 2, est le ratio du carré de la longueur des vecteurs  $X_1\mathbf{e}_1$  et  $X_2\mathbf{e}_2$  sur la figure. Le poids plus avantageux de l'observation 2 signifie que  $X_2$  doit être plus important que  $X_1$ , ainsi que nous l'avons représenté. Au contraire, si  $X_1$  et  $X_2$  étaient sensiblement égaux, de sorte que  $\mathbf{P}_{1,2}\mathbf{X}$  forme approximativement le même angle avec chacun des deux axes, cela traduirait la quasi égalité de  $h_1$  et de  $h_2$ .

Nous prenons maintenant un exemple numérique pour mettre en évidence l'influence que peut avoir une observation erronée isolée. L'exemple montre également que l'examen des points fortement pondérés peut se révéler très utile dans la détection des erreurs qui affectent substantiellement les estimations. L'ensemble de données correct, qui comprend les  $\hat{u}_t$  et les  $h_t$ 's, est détaillé dans le Tableau 1.2. Les estimations OLS correspondantes sont

$$\hat{\beta}_1 = 1.390, \quad \hat{\beta}_2 = 1.223, \quad R^2 = 0.7278. \quad (1.50)$$

Le  $h_t$  le plus fort est ici 0.536, pour l'observation 3. Cette valeur est plus que cinq fois supérieure à la valeur la plus faible des  $h_t$  et elle est plus grande que  $2k/n$  ( $= 0.40$  dans ce cas). Ainsi l'observation 3 est un point fortement

**Tableau 1.2** Exemple Numérique: Bonnes Données

$t$	$\mathbf{X}_t$		$y_t$	$\hat{u}_t$	$h_t$	$\hat{u}_t h_t / (1 - h_t)$
1	1	1.51	2.88	-0.357	0.203	-0.091
2	1	2.33	3.62	-0.620	0.105	-0.073
3	1	3.57	5.64	-0.116	0.536	-0.134
4	1	2.12	3.43	-0.553	0.101	-0.062
5	1	1.54	3.21	-0.064	0.194	-0.015
6	1	1.71	4.49	1.008	0.151	0.179
7	1	2.68	4.50	-0.168	0.156	-0.031
8	1	2.25	4.28	-0.138	0.101	0.016
9	1	1.32	2.98	-0.025	0.269	-0.009
10	1	2.80	5.57	0.755	0.186	0.173

**Tableau 1.3** Exemple Numérique: Mauvaises Données

$t$	$\mathbf{X}_t$		$y_t$	$\hat{u}_t$	$h_t$	$\hat{u}_t h_t / (1 - h_t)$
1	1	1.51	2.88	-0.900	0.143	-0.150
2	1	2.33	3.62	-0.356	0.104	-0.041
3	1	3.57	5.64	1.369	0.125	0.195
4	1	2.12	3.43	-0.496	0.110	-0.061
5	1	1.54	3.21	-0.578	0.141	-0.095
6	1	1.71	4.49	0.662	0.130	0.099
7	1	7.68	4.50	-0.751	0.883	-5.674
8	1	2.25	4.28	0.323	0.106	0.038
9	1	1.32	2.98	-0.755	0.158	-0.142
10	1	2.80	5.57	1.482	0.100	0.165

pondéré. Cela n'a rien d'étonnant puisque la valeur de  $X_{2t}$  pour l'observation 3 est de loin la plus importante des  $X_{2t}$ . Cependant, deux autres observations ont aussi des valeurs de  $h_t$  supérieures à 0.20 de sorte que l'observation 3 n'est pas un point extrêmement pondéré. Comme le montre également la dernière colonne du tableau, elle n'a pas une influence particulièrement forte.

Maintenant, observons ce qu'il adviendrait si nous introduisions délibérément une erreur dans  $\mathbf{X}$ . Supposons que  $X_{2t}$  pour l'observation 7 devienne subitement 7.68. L'ensemble des données correspondantes, ainsi que les  $\hat{u}_t$  et  $h_t$  est reporté dans le Tableau 1.3. Les estimations OLS qui y sont rattachées sont

$$\hat{\beta}_1 = 3.420, \quad \hat{\beta}_2 = 0.238, \quad R^2 = 0.1996.$$

Ces estimations diffèrent considérablement de celles calculées en (1.50). La relation solide qui reliait  $X_{2t}$  et  $y_t$  précédemment a tout simplement disparu consécutivement à l'introduction d'une erreur dans l'observation 7.<sup>11</sup> L'examen des seuls résidus ne nous renseignerait pas qu'il y a quelque chose qui "cloche" avec cette observation, puisque  $\hat{u}_7$  n'est pas le résidu le plus important. Par contre, l'examen des  $h_t$  suggérerait de lire plus attentivement les observations; parce que  $h_7$  est plus que cinq fois plus important que n'importe quel autre  $h_t$ , l'observation 7 est en fait un point de très forte pondération. Mais c'est aussi un point de forte influence, comme le montre la dernière colonne du tableau. Alors dans cette situation quiconque observerait soit  $h_t$  soit  $\hat{u}_t h_t / (1 - h_t)$  serait selon toute vraisemblance amené à déceler la donnée nuisible.

Cet exemple suggère que l'économetre prudent observera les quantités  $h_t$  et  $\hat{u}_t h_t / (1 - h_t)$  comme une formalité. Malheureusement, tous les logiciels de régression n'intègrent pas ces possibilités de calcul. Cela est d'autant plus surprenant que les  $h_t$  sont facilement calculables si les estimations OLS sont exécutées à l'aide de la décomposition QR. Puisque  $\mathbf{P}_X = \mathbf{Q}\mathbf{Q}^\top$ , on aperçoit aisément que

$$h_t = \sum_{i=1}^k Q_{ti}^2$$

ce qui rend le calcul extrêmement simple lorsque l'on possède  $\mathbf{Q}$ . Il est possible, une fois que l'on a calculé les  $h_t$  et/ou les  $\hat{u}_t h_t / (1 - h_t)$ , de les tracer par rapport à  $t$ . S'il y a des points fortement pondérés et/ou des observations d'une influence injustifiée, alors il sera sage de vérifier la précision des données ou de savoir dans quelles proportions leur omission de l'échantillon affecte les résultats. De telles procédures informelles pour la détection des observations influentes, et plus particulièrement celles générées par des erreurs de données, fonctionnent généralement avec succès. Mais des procédures plus formelles sont détaillées par Belsley, Kuh, et Welsch (1980) et Krasker, Kuh, et Welsch (1983), dans les références que nous avons données.

<sup>11</sup> Puisque nous ne discutons que des aspects numériques des moindres carrés dans ce chapitre, nous n'avons pas présenté les écarts types ou les statistiques  $t$  pour cet exemple. Nous observons néanmoins que la différence entre ces deux ensembles d'estimations basés sur les données exactes et inexactes est importante relativement aux écarts types que l'on calcule généralement; par exemple la statistique  $t$  pour le  $\hat{\beta}_2$  correct (1.223) est 4.62, alors que la statistique  $t$  pour  $\hat{\beta}_2$  découlant du mauvais ensemble de données (0.238) est seulement égal à 1.41.

## 1.7 LECTURES COMPLÉMENTAIRES ET CONCLUSION

L'usage de la géométrie comme point d'appui pour la compréhension de la régression linéaire n'est pas un fait récent; consulter Herr (1980). Les articles les plus anciens et les plus importants sont ceux de Fisher (1915), Durbin et Kendall (1951), Kruskal (1961, 1968, 1975), et Seber (1964). Une référence précieuse que l'on peut consulter concernant l'approche géométrique des modèles linéaires est Seber (1980), bien que son ouvrage puisse paraître trop concis pour certains lecteurs. Un article récent et néanmoins accessible est celui de Bryant (1984). L'approche n'a pas été autant usitée en économétrie qu'elle ne l'a été en statistiques, mais un certain nombre de textes et d'articles économétriques — (notamment Malinvaud (1970a) ainsi que Madansky (1976), Pollock (1979), et Wonnacott et Wonnacott (1979) — en font un usage plus ou moins intensif. On pourrait qualifier notre approche de *semi-géométrie*, puisque nous n'avons pas insisté autant que certains auteurs sur la nature sans coordonnées de notre analyse; consulter les articles de Kruskal, l'ouvrage de Seber ou dans un domaine typiquement économétrique, Fisher (1981, 1983) et Fisher et McAleer (1984).

Nous avons totalement fait abstraction des modèles statistiques dans ce chapitre. La régression linéaire a été considérée comme un mécanisme d'estimation qui a une interprétation géométrique, plutôt que comme une procédure d'estimation opérant sur une famille de modèles statistiques. Tous les résultats que nous avons énoncés se sont révélés numériquement exacts, ce qui découle de la façon dont on calcule les estimations des moindres carrés, et ne sont en rien relatifs à la manière dont on a généré les données. Il nous faut insister sur ce point, car les traitements usuels du modèle de régression linéaire font rarement cette distinction entre les propriétés statistiques et les propriétés numériques des moindres carrés.

Nos discussions ultérieures dans cet ouvrage porteront sur l'étude de divers modèles statistiques, dont certains seront des modèles de régression et d'autres non, que les économètres utilisent dans la pratique. Dans la majeure partie du livre, nous porterons notre attention sur deux sortes de modèles: ceux que l'on peut estimer comme des modèles de régression linéaire ou non linéaire, et ceux que l'on doit estimer par la méthode du maximum de vraisemblance (cette dernière classe regroupant en fait un grand nombre de modèles). Comme nous aurons l'occasion de le constater, la compréhension des propriétés géométriques d'une régression linéaire s'avère être fondamentale pour la compréhension des modèles de régression non linéaire et de la méthode du maximum de vraisemblance. Nous supposerons donc dans ce qui suit que les lecteurs sont familiers avec tous les résultats de base que nous avons énoncés au cours de ce chapitre.



## TERMES ET CONCEPTS

base orthogonale	pondération
champ de projection	problème insuffisamment conditionné
codimension (d'un sous-espace linéaire)	projection
coefficient de détermination ( $R^2$ )	propriétés numériques ou propriétés statistiques
complément orthogonal (d'un sous-espace)	régressande
décomposition orthogonale	régresseur
dimension (d'un sous-espace linéaire)	régression, linéaire et non linéaire
équations normales	reparamétrisation (d'une régression)
espace euclidien de dimension $n$ , $E^n$	résidus
$h_t$ (éléments diagonaux de la matrice "chapeau")	résidus des moindres carrés
indépendance linéaire	résultats sans coordonnées
longueur d'un vecteur	$R^2$ , centré et non centré
matrice "chapeau"	somme des carrés expliqués, ESS
matrice de projection orthogonale	somme des carrés totaux, TSS
matrice idempotente	somme des résidus au carré, SSR
modèle équilibré	sous-espace engendré par les colonnes d'une matrice
moindres carrés ordinaires (OLS)	Théorème de Pythagore
moindres carrés contraints	Théorème FWL
nombre à virgule flottante	valeurs ajustées
observations influentes	vecteur unité ou constante, $\mathbf{1}$
	vecteurs orthogonaux

# Chapitre 2

## Modèles de Régression non Linéaire et les Moindres Carrés non Linéaires

### 2.1 INTRODUCTION

Dans le Chapitre 1, nous avons discuté en détail de la géométrie des moindres carrés ordinaires et de leurs propriétés en tant que système de calcul. Ce matériau est important car de nombreux modèles statistiques communément usités sont souvent estimés à l'aide de variantes des moindres carrés. Parmi ceux-ci, nous trouvons le type de modèle le plus communément rencontré en économétrie, c'est-à-dire la classe des **modèles de régression**, dont nous entamons l'étude dès à présent. Au lieu de nous restreindre volontairement au domaine bien connu des **modèles de régression linéaire**, qu'il est possible d'estimer directement par OLS, nous considérons la famille plus large des **modèles de régression non linéaire** qui peuvent être estimés par **moindres carrés non linéaires**, ou **NLS**. Parfois, nous traiterons de manière spécifique des modèles de régression linéaire si les résultats qui sont vérifiés pour de tels modèles ne se généralisent pas au cas non linéaire.

Au cours de ce chapitre et des quelques chapitres suivants consacrés aux modèles de régression, nous porterons notre attention principalement sur les **modèles univariés**, c'est-à-dire les modèles dans lesquels n'existe qu'une seule variable dépendante. Ceux-ci sont beaucoup plus simples à traiter que les **modèles multivariés** dans lesquels on trouve plusieurs variables dépendantes jointes. Les modèles univariés sont de loin plus fréquemment rencontrés en pratique que les modèles multivariés, et une bonne compréhension des premiers est essentielle pour une bonne compréhension des seconds. Nous démontrerons au Chapitre 9 qu'il est simple de rendre compatibles les résultats des modèles univariés aux modèles multivariés.

Nous commençons par écrire le modèle de régression linéaire univariée sous sa forme générique:

$$y_t = x_t(\beta) + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad t = 1, \dots, n. \quad (2.01)$$

Désormais,  $y_t$  représente l'observation  $t$  de la **variable dépendante**, qui est une variable aléatoire scalaire, et  $\beta$  désigne un vecteur à  $k$  composantes de

paramètres (généralement) inconnus. La fonction scalaire  $x_t(\beta)$  est une **fonction de régression** (le plus souvent non linéaire) qui détermine l'espérance de  $y_t$  conditionnelle à  $\beta$  et (généralement) à certaines **variables indépendantes**. Ces dernières n'ont pas été mentionnées explicitement dans (2.01) mais le  $t$  en indice de  $x_t(\beta)$  indique que cette fonction varie d'une observation à l'autre. Dans la plupart des cas, cela s'explique parce que  $x_t(\beta)$  dépend d'une ou plusieurs variables indépendantes qui varient. Ainsi,  $x_t(\beta)$  devrait être interprétée comme l'espérance de  $y_t$  *conditionnelle* aux valeurs de ces variables indépendantes. De manière plus précise, comme nous aurons l'occasion de le voir à la Section 2.4, elle devrait être interprétée comme l'espérance de  $y_t$  conditionnelle à un **ensemble d'informations** auxquels appartiennent ces variables indépendantes.<sup>1</sup>

Dans certains cas,  $x_t(\beta)$  pourra aussi dépendre de variables retardées de  $y_t$ . Un modèle qui comportera une telle fonction de régression sera appelé **modèle dynamique**, et le traitement de ce genre de modèles complique quelque peu l'analyse. Nous admettrons pour l'instant que  $x_t(\beta)$  ne dépend pas des valeurs retardées de  $y_t$ , contrairement à ce qui serait le cas si (2.01) était un modèle dynamique, mais nous abandonnerons cette hypothèse au Chapitre 5 lorsque nous présenterons un premier traitement de la théorie asymptotique des moindres carrés non linéaires. D'après l'acception du terme que nous utilisons dans cet ouvrage, les résultats **asymptotiques** ne sont vrais qu'à la limite, lorsque la taille  $n$  de l'échantillon tend vers l'infini. La plupart des résultats analytiques standards concernant les modèles de régression non linéaire, et les modèles non linéaires en général, sont des résultats asymptotiques, parce que les résultats établis à l'aide d'échantillons finis et faciles à interpréter sont souvent extrêmement difficiles à obtenir.

Les **modèles de régression** se différencient de tous les autres modèles statistiques par le fait que l'aléa affecte les variables dépendantes uniquement par l'intermédiaire d'un **aléa** additif. Dans le cas précis de (2.01), cet aléa est appelé  $u_t$ , et la notation " $u_t \sim \text{IID}(0, \sigma^2)$ " est un moyen concis pour dire que les aléas  $u_t$  sont supposés être **indépendants et identiquement distribués**, ou **i.i.d.**, avec une espérance nulle et une variance égale à  $\sigma^2$ . En prétendant cela, nous ne voulons pas dire que les variables aléatoires  $u_t$  ont nécessairement la même distribution, mais simplement qu'elles sont d'espérance zéro et de

<sup>1</sup> Les lecteurs devraient être avertis que la notation que nous avons utilisée ici est quelque peu inhabituelle. De nombreux auteurs utilisent  $f_t(\beta)$  en lieu et place de notre  $x_t(\beta)$ . Nous préférons cette notation pour deux raisons. La première est qu'elle nous laisse la liberté d'utiliser la notation  $f(\cdot)$  pour désigner des objets autres que les fonctions de régression sans créer d'ambiguïté. La seconde est qu'avec notre notation, il devient naturel de désigner  $\partial x_t(\beta)/\partial \beta_i$  par  $X_{ti}(\beta)$  (voir la Section 2.2). La matrice dont l'élément type est  $X_{ti}(\beta)$  est de fait étroitement liée à la matrice habituelle  $\mathbf{X}$  qui est utilisée dans la plupart des traitements du modèle de régression linéaire, et nous espérons que cette ressemblance d'écriture sera un moyen efficace de se le rappeler.

variance  $\sigma^2$ . A ce propos, les lecteurs devraient sans doute être avertis que nous dérogeons à l'usage standard. Ainsi que nous le verrons dans la Section 2.6, les propriétés de ces aléas sont cruciales car elles déterminent toutes les propriétés statistiques du modèle, et par là, permettent de savoir si un modèle de régression peut raisonnablement être utilisé ou pas. Quoi qu'il en soit, puisque les estimations NLS (comme les estimations OLS) peuvent être calculées sans se préoccuper de la façon dont les données ont été générées, nous traiterons le calcul des estimations NLS avant d'aborder la discussion de leurs propriétés statistiques.

Le reste du chapitre traite un certain nombre d'aspects des moindres carrés non linéaires et des modèles de régression non linéaire. Dans la Section 2.2, nous discutons des moindres carrés non linéaires en tant que procédure de calcul qui constitue une extension des moindres carrés ordinaires. Nous démontrons que la minimisation de la somme des résidus au carré pour un modèle de régression non linéaire tel que (2.01) est très semblable, eu égard à la géométrie impliquée, à l'exécution d'une régression linéaire. Un modèle de régression non linéaire doit être **identifié** si l'on désire obtenir des estimations uniques des paramètres. Nous discutons par conséquent du concept fondamental d'identification dans la Section 2.3. Dans la seconde moitié du présent chapitre, nous entamerons la discussion des aspects statistiques (et économiques) des modèles de régression non linéaire. Dans la Section 2.4 nous verrons comment les équations de régression comme (2.01) s'interprètent, et la distinction entre les modèles et les processus générateurs de données. Puis des exemples de fonctions de régression linéaires et non linéaires seront examinés à la Section 2.5, alors que les aléas seront examinés à la Section 2.6. Procéder à des inférences à partir de modèles estimés par NLS sera le thème du Chapitre 3.

## 2.2 LA GÉOMÉTRIE DES MOINDRES CARRÉS NON LINÉAIRES

Le moyen de loin le plus répandu d'estimer aussi bien les modèles de régression non linéaire que les modèles de régression linéaire, consiste à minimiser la somme des résidus au carré, ou SSR, en tant que fonction de  $\beta$ . En ce qui concerne le modèle (2.01), la **fonction somme-des-carrés** est

$$SSR(\beta) = \sum_{t=1}^n (y_t - x_t(\beta))^2.$$

L'écriture de cette expression sous forme matricielle est généralement plus pratique:

$$SSR(\beta) = (\mathbf{y} - \mathbf{x}(\beta))^\top (\mathbf{y} - \mathbf{x}(\beta)), \quad (2.02)$$

où  $\mathbf{y}$  désigne un vecteur à  $n$  composantes d'observations  $y_t$ , et  $\mathbf{x}(\beta)$  représente un vecteur composé de  $n$  fonctions de régression  $x_t(\beta)$ . Ainsi que nous l'avons

constaté à l'occasion du Chapitre 1, une notation alternative, qui n'apparaît peut-être pas aussi facile à manipuler algébriquement, mais qui est plus concise, met l'accent sur l'aspect géométrique,

$$SSR(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})\|^2, \quad (2.03)$$

où  $\|\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})\|$  mesure la longueur du vecteur  $\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})$ . Selon l'expression (2.03) il est clair que lorsque l'on minimise  $SSR(\boldsymbol{\beta})$ , on minimise en fait la distance euclidienne entre  $\mathbf{y}$  et  $\mathbf{x}(\boldsymbol{\beta})$ , dont nous discuterons plus longuement de l'interprétation plus loin.

La fonction somme-des-carrés (2.02) peut être réécrite comme

$$SSR(\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{x}(\boldsymbol{\beta}) + \mathbf{x}^\top(\boldsymbol{\beta}) \mathbf{x}(\boldsymbol{\beta}).$$

En dérivant cette expression par rapport à toutes les composantes du vecteur  $\boldsymbol{\beta}$  à  $k$  éléments, et en annulant toutes les dérivées partielles, nous obtenons les conditions du premier ordre qui doivent être vérifiées pour toute estimation NLS du vecteur  $\hat{\boldsymbol{\beta}}$  qui correspond à un minimum intérieur de  $SSR(\boldsymbol{\beta})$ . Ces conditions du premier ordre, ou équations normales, sont

$$-2\mathbf{X}^\top(\hat{\boldsymbol{\beta}})\mathbf{y} + 2\mathbf{X}^\top(\hat{\boldsymbol{\beta}})\mathbf{x}(\hat{\boldsymbol{\beta}}) = \mathbf{0}, \quad (2.04)$$

où la matrice  $\mathbf{X}(\boldsymbol{\beta})$  de dimension  $n \times k$  est composée d'éléments tels que

$$X_{ti}(\boldsymbol{\beta}) \equiv \frac{\partial x_t(\boldsymbol{\beta})}{\partial \beta_i}.$$

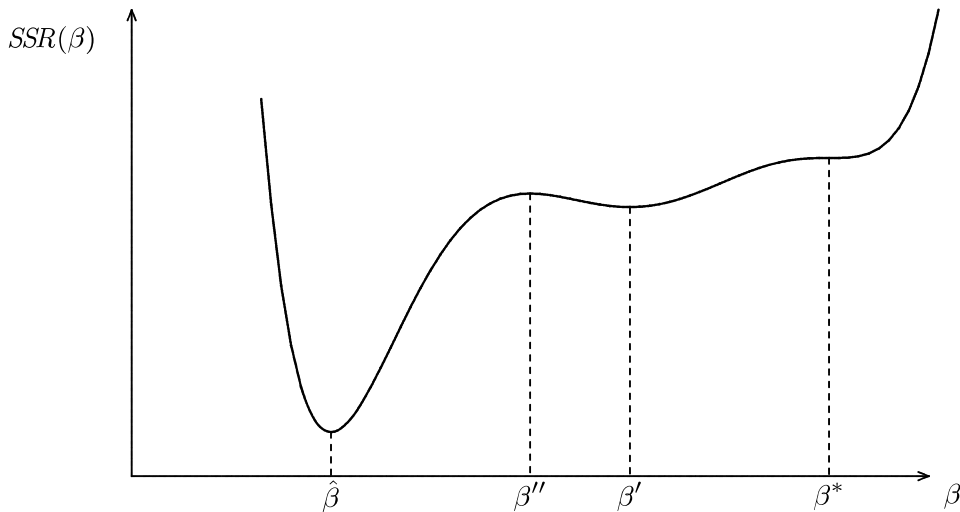
Le fait que chaque vecteur de (2.04) possède  $k$  éléments implique l'existence de  $k$  équations normales déterminant les  $k$  composantes de  $\boldsymbol{\beta}$ .

Nous retrouverons à plusieurs reprises la matrice  $\mathbf{X}(\boldsymbol{\beta})$  lors de notre discussion sur les moindres carrés non linéaires. Chaque élément de cette matrice correspond à la dérivée partielle d'un élément de  $\mathbf{x}(\boldsymbol{\beta})$  par rapport à un élément de  $\boldsymbol{\beta}$ . Comme la notation que nous avons adoptée le suggère, la matrice  $\mathbf{X}(\boldsymbol{\beta})$  correspond exactement à la matrice  $\mathbf{X}$  dans le cas de la régression linéaire. Ainsi, lorsque la fonction de régression  $\mathbf{x}(\boldsymbol{\beta})$  s'apparente à la fonction linéaire  $\mathbf{X}\boldsymbol{\beta}$ , nous voyons immédiatement que  $\mathbf{X}(\boldsymbol{\beta}) = \mathbf{X}$ .

Les conditions du premier ordre (2.04) peuvent légèrement se simplifier en regroupant les termes, en éliminant le facteur  $-2$ , et en adoptant les définitions  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\boldsymbol{\beta}})$  et  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\boldsymbol{\beta}})$ .<sup>2</sup> Le résultat est

$$\hat{\mathbf{X}}^\top(\mathbf{y} - \hat{\mathbf{x}}) = \mathbf{0}. \quad (2.05)$$

<sup>2</sup> Il est souvent pratique d'indiquer de cette façon la dépendance d'un vecteur ou d'une matrice par rapport à un vecteur de paramètres qui a été estimé. Ainsi, si  $\boldsymbol{\alpha}_0$  était un ensemble de paramètres exact, et  $\hat{\boldsymbol{\alpha}}$  et  $\tilde{\boldsymbol{\alpha}}$  deux ensembles d'estimations, alors  $\mathbf{Z}_0$  désignerait  $\mathbf{Z}(\boldsymbol{\alpha}_0)$ ,  $\hat{\mathbf{Z}}$  désignerait  $\mathbf{Z}(\hat{\boldsymbol{\alpha}})$ , et  $\tilde{\mathbf{Z}}$  désignerait  $\mathbf{Z}(\tilde{\boldsymbol{\alpha}})$ .



**Figure 2.1** Une fonction somme des carrés

Ces équations normales nous enseignent simplement que les résidus  $\mathbf{y} - \hat{\mathbf{x}}$  doivent être orthogonaux à la matrice des dérivées  $\hat{\mathbf{X}}$ . Il s'agit d'un résultat analogue à celui obtenu pour les modèles de régression linéaire pour lesquels les résidus  $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$  doivent être orthogonaux à la matrice  $\mathbf{X}$ . La différence entre les cas linéaire et non linéaire réside dans le fait qu'autant le vecteur de valeurs ajustées  $\hat{\mathbf{x}}$  que la matrice  $\hat{\mathbf{X}}$  dépendent de  $\hat{\boldsymbol{\beta}}$ . Ainsi en général, nous ne pouvons pas espérer résoudre (2.05) analytiquement pour  $\hat{\boldsymbol{\beta}}$ , bien que cela soit réalisable dans certains cas particuliers, dont bien sûr le cas linéaire.

Notons que les conditions du premier ordre (2.05) sont nécessaires mais non suffisantes pour faire de  $\hat{\boldsymbol{\beta}}$  un minimum intérieur et global de la fonction somme des carrés. Il peut exister plusieurs valeurs de  $\boldsymbol{\beta}$  qui vérifient (2.05) et qui correspondent à des **minima locaux**, des **points stationnaires** et même des **maxima locaux**. Cela est illustré sur la Figure 2.1 pour le cas où il n'y a qu'un seul paramètre, faisant de  $\beta$  un scalaire. Sur la figure, le minimum global se situe en  $\hat{\beta}$ , mais apparaissent également un autre minimum local en  $\beta'$ , un maximum local en  $\beta''$ , et un point stationnaire en  $\beta^*$ .

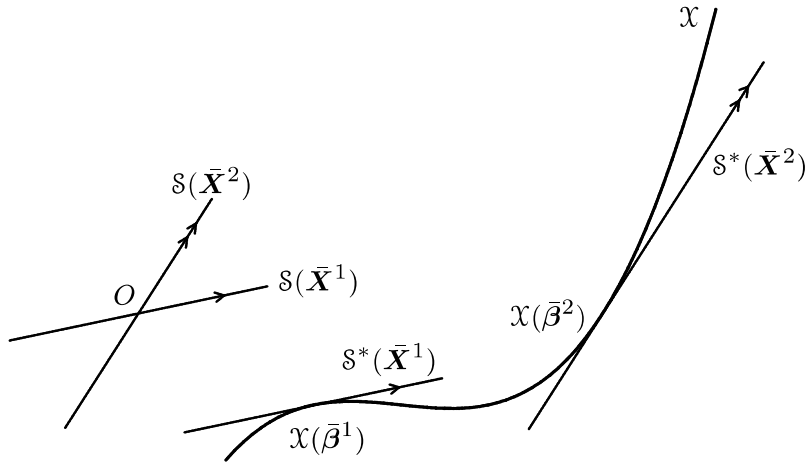
Aucun **algorithme de minimisation** efficace ne s'arrêtera sur un maximum local ou un point stationnaire, parce qu'il est aisé de vérifier que les conditions du second ordre ne seraient pas satisfaites pour de tels points. Mais un algorithme pourra ne pas déceler un minimum global et s'arrêter à un minimum local. En se basant uniquement sur des informations locales, aucun algorithme ne distingue un minimum local comme  $\beta'$  d'un minimum global comme  $\hat{\beta}$ . Dans le but de trouver le minimum global, il est donc nécessaire de minimiser  $SSR(\boldsymbol{\beta})$  un certain nombre de fois, en débutant par une variété de points de départ différents. Dans l'exemple que nous avons illustré, un algorithme efficace serait capable de trouver  $\hat{\beta}$  seulement s'il débute à partir d'un point quelconque situé à gauche de  $\beta''$ . Dans le cas unidimensionnel, il

est aisé de trouver avec certitude un minimum global, dès lors qu'un graphe similaire à la Figure 2.1 permet de le repérer. Cependant, dans le cas où le nombre de dimensions est plus élevé, les méthodes graphiques ne sont en général d'aucune utilité, et même lorsque l'on démarre un algorithme avec un certain nombre de points de départ, il n'existe aucune garantie de trouver le minimum global si l'on obtient plusieurs minima locaux. Des méthodes de calcul des estimations NLS seront discutées plus tard, au Chapitre 6.

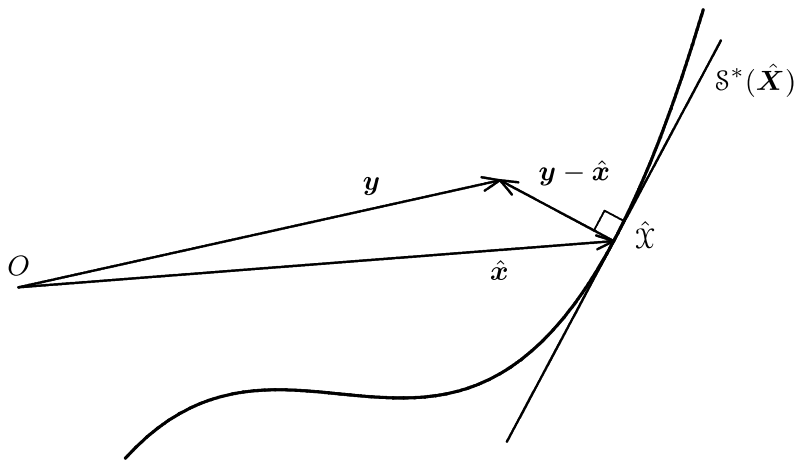
Il est instructif d'étudier l'analogie des Figures 1.1 et 1.3 pour le cas de la régression non linéaire. Souvenons-nous que  $\mathbf{y}$  peut être considéré comme un point dans l'espace des observations  $E^n$ , et que la fonction de régression linéaire  $\mathbf{X}\boldsymbol{\beta}$  définit alors un sous-espace à  $k$  dimensions de cet espace. Dans la Figure 1.3, nous avons illustré, pour le cas le plus simple où  $n = 2$  et  $k = 1$ , la façon dont les moindres carrés ordinaires projettent  $\mathbf{y}$  orthogonalement sur  $\mathcal{S}(\mathbf{X})$ , le sous-espace engendré par les colonnes de  $\mathbf{X}$ . Lorsque la fonction de régression  $\mathbf{x}(\boldsymbol{\beta})$  est non linéaire, mais partout différentiable, elle définit une **variété** à  $k$  dimensions,<sup>3</sup> ou une surface lisse, qui ne constitue plus un sous-espace *linéaire* en général. Chaque point de cette variété, que nous noterons  $\mathcal{X}$ , correspond (par hypothèse) à une valeur différente de  $\boldsymbol{\beta}$ , et donc on pourra se référer à un point particulier qui correspond à  $\boldsymbol{\beta}^1$ , en le notant  $\mathcal{X}(\boldsymbol{\beta}^1)$ . Il est essentiel pour que  $\mathcal{X}$  soit lisse partout, que chaque composante du vecteur  $\mathcal{S}(\boldsymbol{\beta})$  soit partout dérivable. Pour n'importe quel point choisi arbitrairement, disons  $\bar{\boldsymbol{\beta}}$ , la matrice  $\bar{\mathbf{X}} \equiv \mathbf{X}(\bar{\boldsymbol{\beta}})$  définit un **espace tangent**  $\mathcal{S}^*(\bar{\mathbf{X}})$ , qui correspond tout simplement au sous-espace linéaire à  $k$  dimensions  $\mathcal{S}(\bar{\mathbf{X}})$ , translaté de façon à avoir l'origine en  $\mathcal{X}(\bar{\boldsymbol{\beta}})$ . Cela implique que  $\mathcal{S}^*(\bar{\mathbf{X}})$  est tangent à  $\mathcal{X}$  en ce point.

La Figure 2.2 illustre ces considérations dans le cas  $k = 1$ . On suppose que  $\mathbf{x}(\boldsymbol{\beta})$  se situe, au moins localement, dans un sous-espace de  $\mathbb{R}^n$  à deux dimensions, ce qui nous permet de le dessiner sur la feuille. La figure représente la variété incurvée  $\mathcal{X}$ , les espaces tangents  $\mathcal{S}^*(\bar{\mathbf{X}}^1)$  et  $\mathcal{S}^*(\bar{\mathbf{X}}^2)$  en deux points arbitrairement choisis  $\mathcal{X}(\bar{\boldsymbol{\beta}}^1)$  et  $\mathcal{X}(\bar{\boldsymbol{\beta}}^2)$ , et les sous-espaces linéaires correspondants  $\mathcal{S}(\bar{\mathbf{X}}^1)$  et  $\mathcal{S}(\bar{\mathbf{X}}^2)$ . Ces derniers, comme les flèches sur la figure l'indiquent, sont parallèles à  $\mathcal{S}^*(\bar{\mathbf{X}}^1)$  et  $\mathcal{S}^*(\bar{\mathbf{X}}^2)$  respectivement, mais ne sont pas mutuellement parallèles. Si  $\mathcal{X}$  était rectiligne, comme cela serait le cas si la fonction de régression était linéaire, alors bien évidemment il n'y aurait pas de distinction possible entre  $\mathcal{X}$ ,  $\mathcal{S}(\bar{\mathbf{X}}^1)$ ,  $\mathcal{S}(\bar{\mathbf{X}}^2)$ ,  $\mathcal{S}^*(\bar{\mathbf{X}}^1)$ , et  $\mathcal{S}^*(\bar{\mathbf{X}}^2)$ . C'est justement la présence de telles distinctions qui rend les modèles non linéaires plus difficiles à traiter que les modèles linéaires. Notons également que bien que la variété définie par une fonction de régression linéaire comprenne toujours l'origine, ce n'est en général pas le cas pour une fonction non linéaire, comme on peut le constater sur la figure.

<sup>3</sup> Pour des définitions plus formelles d'une variété, ainsi que pour une discussion minutieuse des propriétés des variétés, consulter entre autres, Spivak (1965) pour une approche rudimentaire et Lang (1972) pour une approche plus avancée.



**Figure 2.2** Espaces tangents à une variété incurvée



**Figure 2.3** Une régressande  $y$  projetée sur une variété non linéaire

La Figure 2.3 montre la même variété  $\mathcal{X}$  que la Figure 2.2, mais  $\mathcal{S}(\bar{\mathbf{X}}^1)$ ,  $\mathcal{S}(\bar{\mathbf{X}}^2)$ ,  $\mathcal{S}^*(\bar{\mathbf{X}}^1)$ , et  $\mathcal{S}^*(\bar{\mathbf{X}}^2)$  n'y figurent plus. Apparaissent par contre une régressande  $y$  et sa projection orthogonale sur  $\mathcal{X}$  au point  $\hat{\mathbf{X}} \equiv \mathcal{X}(\hat{\beta})$ . Notons que puisque  $\mathcal{S}^*(\hat{\mathbf{X}})$  est tangent à  $\mathcal{X}$  en  $\hat{\beta}$ ,  $y - \hat{x}$  doit être orthogonal à  $\mathcal{S}^*(\hat{\mathbf{X}})$  ainsi qu'à  $\mathcal{X}$  au point  $\hat{\mathbf{X}}$ , ce que réclament précisément les conditions du premier ordre. Comme sur cette figure la fonction de régression  $x(\beta)$ , et par conséquent la variété  $\mathcal{X}$ , est légèrement non linéaire, n'y a qu'un seul point  $\hat{\mathbf{X}}$  pour lequel les conditions du premier ordre sont satisfaites. Il est clair d'après la figure que  $y$  ne peut être projetée orthogonalement sur  $\mathcal{X}$  qu'en  $\hat{\mathbf{X}}$  et en aucun autre point.

Par contraste, examinons la Figure 2.4. Sur cette figure, la variété est hautement non linéaire, et nous obtenons trois points  $\hat{\mathbf{X}}$ ,  $\mathbf{X}'$ , et  $\mathbf{X}''$  (correspon-



nant respectivement à  $\hat{\beta}$ ,  $\beta'$ , et  $\beta''$ ), pour lesquels les conditions du premier ordre sont satisfaites. Pour chacun de ces trois points, que l'on exprime sous forme générique par la notation  $\bar{X}$ ,  $y - \bar{x}$  forme un angle droit avec  $\bar{X}$ , et donc aussi avec  $S^*(\bar{X})$ . Quoi qu'il en soit, dans ce cas,  $\hat{X}$  correspond à l'évidence à un minimum global,  $X''$  à un minimum local, et  $X'$  à un maximum local de  $SSR(\beta)$ . Ainsi, nous avons une occasion supplémentaire de constater que lorsqu'un point satisfait les conditions du premier ordre, il ne correspond pas pour autant à une estimation NLS.

Il ne fait aucun doute d'après ces figures que le degré de non linéarité de la fonction de régression  $x(\beta)$  est crucial. Lorsque  $x(\beta)$  est quasiment linéaire, les moindres carrés non linéaires sont très similaires aux moindres carrés ordinaires. Lorsqu'au contraire,  $x(\beta)$  revêt un caractère non linéaire très marqué, toutes sortes de phénomènes étranges peuvent survenir. La Figure 2.4 fait simplement allusion à cette dernière remarque, puisqu'il y a plusieurs façons différentes pour des valeurs multiples de  $\beta$  de satisfaire les conditions du premier ordre (2.05) lorsque  $X$  correspond à une variété hautement non linéaire.

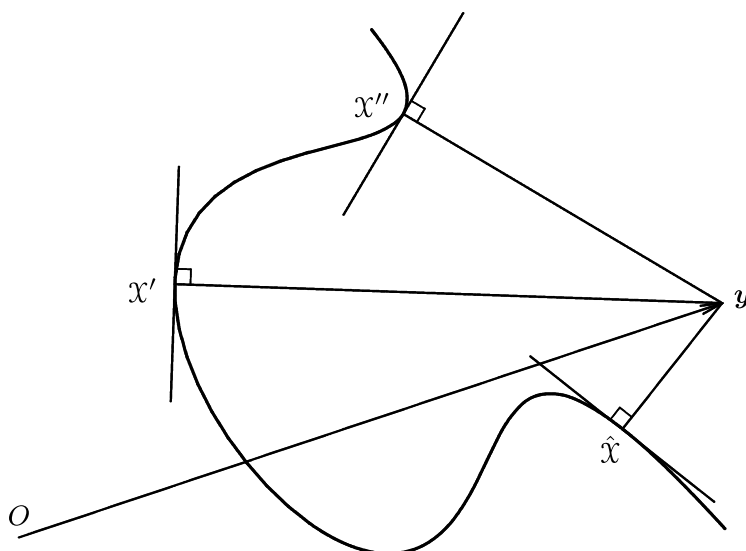
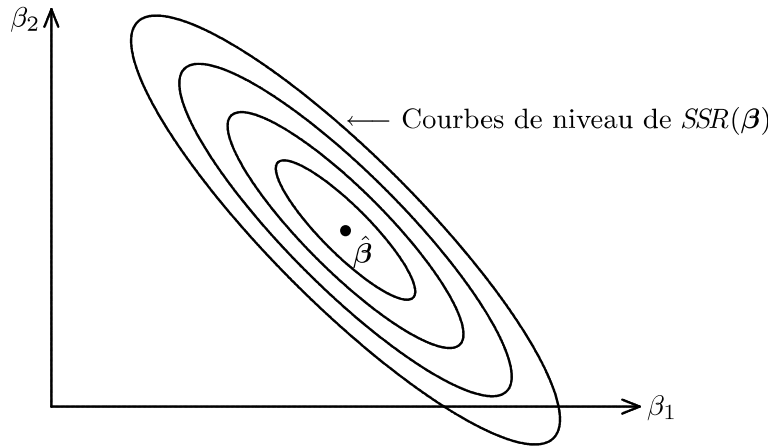


Figure 2.4 Une variété hautement non linéaire

## 2.3 IDENTIFICATION DANS LES MODÈLES NON LINÉAIRES

Pour réussir pleinement la minimisation de  $SSR(\beta)$ , il est nécessaire d'avoir un modèle identifié. **L'identification** évoque un concept géométrique simple qui s'applique à une variété très large de modèles et de techniques d'estimation. Malheureusement, le terme *identification* a été associé dans l'esprit de plusieurs étudiants en économétrie à l'algèbre fastidieuse du modèle d'équations linéaires simultanées. L'identification est en fait un résultat pour de tels modèles, et il existe quelques problèmes particuliers qui apparaissent



**Figure 2.5** Minimum identifié d'une fonction somme des carrés

à leur sujet (consulter les Chapitres 7 et 18), mais il s'agit un concept qui s'applique à *tout* modèle économétrique. Pour l'essentiel, un modèle de moindres carrés non linéaires est **identifié par un ensemble d'informations** donné si, pour cet ensemble de données, il est possible de trouver un  $\hat{\beta}$  *unique* qui minimise  $SSR(\beta)$ . Si le modèle n'est pas identifié par les données utilisées, il existera plus d'un  $\hat{\beta}$ , et peut-être un nombre infini d'entre eux. Certains modèles peuvent n'être identifiés par aucun ensemble concevable de données, alors que d'autres peuvent être identifiés par quelques ensembles de données, mais pas par tous.

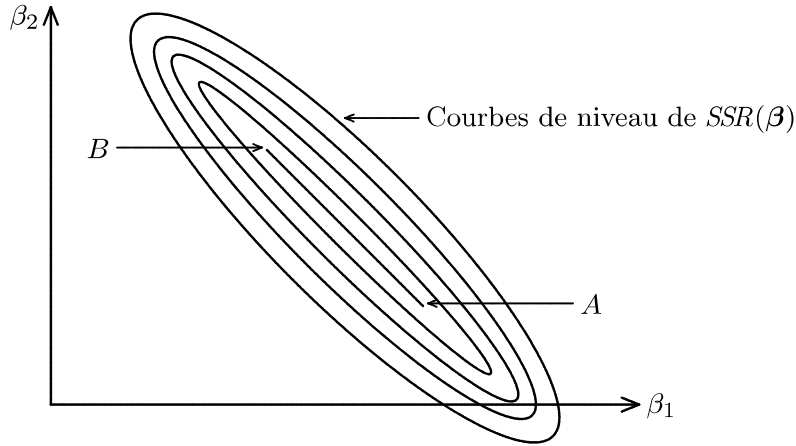
On distingue deux sortes d'identifications, l'identification **locale** et l'identification **globale**. Les estimations  $\hat{\beta}$  des moindres carrés seront **identifiées localement** si pour toute modification légère de  $\hat{\beta}$ , la valeur de  $SSR(\beta)$  s'élève. On peut établir cette définition de façon formelle comme la nécessité d'avoir une fonction  $SSR(\beta)$  strictement convexe en  $\hat{\beta}$ , de sorte que

$$SSR(\hat{\beta}) < SSR(\hat{\beta} + \delta)$$

pour une “petite” variation  $\delta$ . Souvenons-nous que la convexité stricte est vérifiée si la matrice Hessienne  $H(\beta)$ , dont l'élément type est

$$H_{ij}(\beta) \equiv \frac{\partial^2 SSR(\beta)}{\partial \beta_i \partial \beta_j},$$

est définie positive en  $\hat{\beta}$ . La stricte convexité implique que  $SSR(\beta)$  soit incurvée dans toutes les directions; aucun plat n'est autorisé quelle que soit la direction. Si  $SSR(\beta)$  était plate dans une direction au voisinage de  $\hat{\beta}$ , il serait possible de s'éloigner de  $\hat{\beta}$  dans cette direction sans jamais modifier la valeur de la somme des résidus au carré (rappelons-nous que les dérivées premières de  $SSR(\beta)$  sont nulles en  $\hat{\beta}$ , de sorte que  $SSR(\beta)$  doit être égale à  $SSR(\hat{\beta})$  en tout



**Figure 2.6** Minimum non identifié d'une fonction somme des carrés

point de cette région). Par conséquent  $\hat{\beta}$  ne serait pas l'unique estimateur NLS, mais au plus un des points parmi le nombre infini de ceux qui minimisent tous  $SSR(\beta)$ . La Figure 2.5 illustre les courbes de niveau de  $SSR(\beta)$  pour le cas habituel où  $\hat{\beta}$  correspond à un minimum local unique, alors que la Figure 2.6 les représente pour le cas où le modèle n'est pas identifié, parce que tous les points le long de la ligne  $AB$  minimisent  $SSR(\beta)$ .

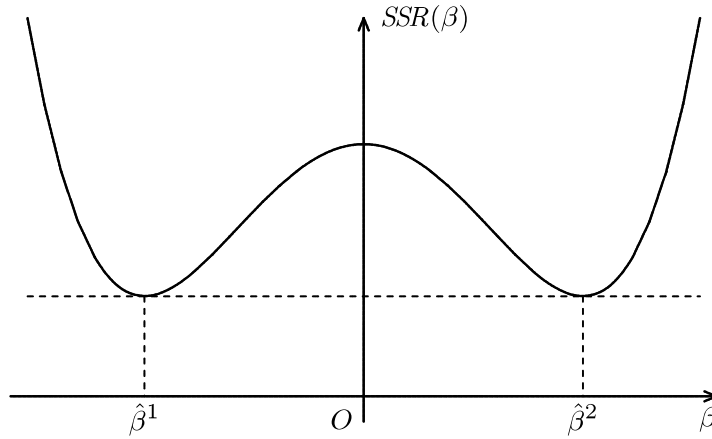
L'identification locale est nécessaire mais non suffisante pour nous fournir une estimation  $\hat{\beta}$  unique. Une condition plus générale est **l'identification globale**, que l'on établit formellement par

$$SSR(\hat{\beta}) < SSR(\beta^*) \quad \text{pour tout} \quad \beta^* \neq \hat{\beta}.$$

Cette définition de l'identification globale reste, à vrai dire, une simple reformulation de la condition d'obtenir un  $\hat{\beta}$  *unique* qui minimise  $SSR(\hat{\beta})$ . Remarquons que même si un modèle est identifié localement, il est toujours possible qu'il y ait deux (ou davantage) estimations distinctes, soit  $\hat{\beta}^1$  et  $\hat{\beta}^2$ , avec  $SSR(\hat{\beta}^1) = SSR(\hat{\beta}^2)$ . A titre d'exemple, examinons le modèle

$$y_t = \beta\gamma + \gamma^2 z_t + u_t. \quad (2.06)$$

Il apparaît clairement que si  $(\hat{\beta}, \hat{\gamma})$  minimise la SSR pour ce modèle,  $(-\hat{\beta}, -\hat{\gamma})$  en fera autant. Donc le modèle est globalement non identifié par *quelque* ensemble de données que ce soit, bien que les conditions du premier ordre et du second ordre soient satisfaites aux deux minima. Cet exemple peut paraître simpliste à première vue, mais le même phénomène apparaît souvent dans de nombreux modèles utilisés par les économistes. Un exemple se trouve être celui des modèles de séries temporelles avec une composante d'erreur à moyenne mobile; consulter le Chapitre 10.



**Figure 2.7** Cas où  $\beta$  est localement identifié mais non globalement

La Figure 2.7 illustre ce que peut donner la fonction somme des carrés pour un modèle qui est localement mais non globalement identifié dans le sens donné précédemment. La fonction somme-des-carrés ne possède qu'un seul argument,  $\beta$ , et elle est symétrique par rapport à l'origine de  $\beta$ . Le minimum de SSR est donc atteint en  $\beta^1$  et en  $\beta^2$ . Chacune des estimations potentielles est identifiée localement, mais le modèle n'est pas identifié globalement.

Il est aussi envisageable d'avoir un modèle globalement identifié, sans pour autant que la condition d'identification locale, impliquant que la matrice Hessienne est définie positive, soit satisfaite, pour certaines valeurs particulières de  $\hat{\beta}$ . Ce genre de lacune d'identification ne pose pas de difficulté si la valeur réalisée  $\hat{\beta}$  se situe assez loin de ces valeurs particulières, et nous parvenons à la calculer, mais il rend difficile l'estimation du modèle. A titre d'exemple, considérons la fonction de régression

$$x_t(\beta) = \beta_1 + \beta_2 z_t^{\beta_3}. \quad (2.07)$$

Il est évident qu'un modèle incorporant cette fonction de régression ne sera pas identifié lorsque  $\hat{\beta}_2 = 0$ , car  $\beta_3$  n'aura alors aucun effet sur la valeur de  $x_t(\beta)$  et par là, aucun effet sur  $SSR(\beta)$ . En conséquence, n'importe quelle valeur de  $\beta_3$  conviendrait pour  $\hat{\beta}_3$ . De façon similaire, le modèle sera non identifié si  $\hat{\beta}_3 = 0$ , car alors  $z_t^{\beta_3}$  et la constante ne pourront être distinguées. Mais parce que  $\hat{\beta}_2$  ou  $\hat{\beta}_3$  ne seront nulles que pour des ensembles de données peu communs, ce modèle sera en réalité identifié par tous les ensembles de données, exception faite de ces ensembles inhabituels.

La fonction de régression (2.07) sert à illustrer un phénomène qu'il est plus fréquent de rencontrer en pratique que les modèles non identifiés, c'est-à-dire des modèles qui sont **insuffisamment identifiés**. Un modèle insuffisamment identifié correspond à un modèle pour lequel la matrice Hessienne  $H(\beta)$  n'est pas véritablement singulière, mais qui devient presque singulière pour des valeurs de  $\beta$  proches de  $\hat{\beta}$ . Ces valeurs de  $\beta$  sont celles qui nous préoccupent

le plus, puisque l'algorithme de minimisation les rencontrera lorsqu'il essaiera de minimiser  $SSR(\beta)$ . Bien que  $SSR(\beta)$  ne soit pas réellement plate pour un modèle insuffisamment identifié, elle est quasiment plate, et ceci pourrait causer quelques problèmes à l'algorithme avec lequel on tente de minimiser  $SSR(\beta)$ . Dans le contexte des modèles de régression linéaire, ce phénomène correspond à la **colinéarité** ou **multicolinéarité** (bien que le préfixe du second terme soit redondant), et il se révèle en rendant la matrice  $\mathbf{X}^\top \mathbf{X}$  presque singulière.

La continuité de la fonction de régression implique qu'un modèle qui incorpore la fonction de régression (2.07) sera insuffisamment identifié s'il arrive que la vraie valeur de  $\beta_2$  ou de  $\beta_3$  soit assez proche de zéro, mais pas véritablement égale. En réalité, il y a de fortes chances pour qu'il soit mal identifié même pour des valeurs de ces paramètres très différentes de zéro, car pour la grande majorité des ensembles de données de  $z_t$ , la Hessienne de ce modèle sera presque singulière. Ainsi que nous le démontrerons au Chapitre 5, la Hessienne  $\mathbf{H}(\beta)$  pour les modèles de régression non linéaire, pour des valeurs  $\beta$  proches de  $\hat{\beta}$ , est généralement assez bien approximée par la matrice

$$2\mathbf{X}^\top(\beta)\mathbf{X}(\beta).$$

Pour la fonction de régression (2.07), la ligne  $t$  de la matrice  $\mathbf{X}(\beta)$  est

$$\begin{bmatrix} 1 & z_t^{\beta_3} & \beta_2 z_t^{\beta_3} \log(z_t) \end{bmatrix}.$$

La troisième colonne de  $\mathbf{X}(\beta)$  est ainsi similaire à la deuxième, chaque élément de celle-ci multiplié par une constante et  $\log(z_t)$  étant égal à l'élément correspondant de la troisième colonne. A moins que l'étendue des valeurs de  $z_t$  ne soit très grande, ou qu'il y ait quelques valeurs de  $z_t$  très proches de zéro,  $z_t^{\beta_3}$  et  $\beta_2 z_t^{\beta_3} \log(z_t)$  tendront à être fortement corrélées, rendant la matrice  $\mathbf{X}^\top(\beta)\mathbf{X}(\beta)$ , et par là la Hessienne dans la plupart des cas, presque singulière. Cet exemple sera examiné en détail dans le Chapitre 6.

Les concepts d'identification locale et globale dont nous venons de discuter diffèrent quelque peu des concepts correspondants d'**identification asymptotique**, que nous verrons au Chapitre 5. Un modèle est identifié asymptotiquement aussi bien localement que globalement si, lorsque la taille  $n$  de l'échantillon tend vers l'infini, le modèle est toujours identifié selon la signification que nous avons donnée. Il s'agit davantage d'une propriété du modèle et de la façon dont les données ont été générées (consulter la Section 2.4 pour une discussion sur les processus générateurs de données) qu'une propriété du modèle et d'un ensemble de données. Comme nous le verrons au cours du Chapitre 5, il est fort possible d'avoir un modèle identifié avec des échantillons finis d'à peu près n'importe quel ensemble de données et pourtant non identifié asymptotiquement; et il est tout aussi envisageable d'avoir un modèle identifié asymptotiquement et non identifié par les nombreux ensembles de données dont on dispose.

## 2.4 MODÈLES ET PROCESSUS GÉNÉRATEURS DE DONNÉES

En économie, rares sont les situations où une relation telle que (2.01) représente réellement la façon dont la variable dépendante est générée, telle qu'elle le serait si  $x_t(\beta)$  était une fonction de réponse à un phénomène physique, et  $u_t$  les erreurs de mesure de  $y_t$ . Au lieu de cela, elle correspond souvent à une façon de modéliser les variations de  $y_t$  causées par les valeurs de certaines variables. Celles-ci peuvent être les seules variables qui soient renseignées, ou celles qui nous intéressent pour un usage particulier. Si nous disposions de davantage d'informations sur les variables explicatives potentielles, nous pourrions fort bien spécifier des  $x_t(\beta)$  différentes en utilisant l'information additionnelle.

Il est quelquefois souhaitable de rendre explicite le fait que  $x_t(\beta)$  représente l'**espérance conditionnelle** de  $y_t$ , c'est-à-dire l'espérance de  $y_t$  dépendant des valeurs d'une quantité d'autres variables. On appelle souvent l'ensemble des variables qui conditionne  $y_t$  l'**ensemble d'informations**. Si l'on note  $\Omega_t$  l'ensemble d'informations qui conditionne la valeur attendue de  $y_t$ , on pourrait définir  $x_t(\beta)$  formellement par  $E(y_t | \Omega_t)$ . Il est possible d'avoir plus d'un ensemble d'informations de ce genre, et donc simultanément

$$x_{1t}(\beta_1) \equiv E(y_t | \Omega_{1t}) \quad \text{et} \quad x_{2t}(\beta_2) \equiv E(y_t | \Omega_{2t}),$$

où  $\Omega_{1t}$  et  $\Omega_{2t}$  représentent les deux ensembles d'informations. Les fonctions  $x_{1t}(\beta_1)$  et  $x_{2t}(\beta_1)$  peuvent différer fortement, et on pourrait vouloir les estimer ensemble à des fins différentes. Il existe plusieurs circonstances pour lesquelles on ne désire pas faire dépendre  $y_t$  de toutes les informations disponibles. Par exemple, si l'on spécifie une fonction de régression dans le but ultime de réaliser des prévisions, il n'y a pas de raison de faire dépendre  $y_t$  des informations qui ne sont pas disponibles pour la période pour laquelle on effectue la prévision. Même lorsque l'on désire intégrer toutes les informations disponibles, le fait qu'une variable particulière appartienne à  $\Omega_t$  n'implique pas qu'elle apparaîtra dans  $x_t(\beta)$ , dès lors que sa valeur ne nous renseigne pas sur l'espérance conditionnelle de  $y_t$ , et l'introduire peut amoindrir notre capacité à estimer l'impact des autres variables sur cette espérance conditionnelle.

Pour toute variable dépendante  $y_t$  donnée et tout ensemble d'informations  $\Omega_t$ , il est toujours possible d'interpréter la différence  $y_t - E(y_t | \Omega_t)$  comme l'aléa associé à l'observation  $t$ . Mais pour qu'un *modèle de régression* soit opérationnel, ces différences doivent généralement avoir la propriété d'être i.i.d.. En fait, il est envisageable, lorsque la taille de l'échantillon est importante, de traiter des cas où les aléas sont indépendants, identiquement distribués uniquement à l'égard des espérances, mais pas forcément à l'égard des variances. Nous discuterons des techniques de traitement de tels cas dans les Chapitres 16 et 17, et dans ce dernier nous abandonnerons l'hypothèse d'indépendance. Comme nous le découvrirons au Chapitre 3 cependant, les techniques conventionnelles pour pratiquer des inférences à partir des

modèles de régression sont sujettes à caution lorsque la propriété d'i.i.d. fait défaut aux modèles, même lorsque la fonction de régression  $x_t(\beta)$  est “correctement” spécifiée. Ainsi, nous perdons toute liberté dans le choix arbitraire de l'ensemble d'informations et dans l'estimation d'une fonction de régression définie et basée sur cet ensemble lorsque nous désirons procéder à des inférences à partir des procédures conventionnelles.

Il existe malgré tout des cas exceptionnels pour lesquels on peut choisir n'importe quel ensemble d'informations, car les modèles établis sur les différents ensembles d'informations seront toujours mutuellement cohérents. Par exemple, supposons que le vecteur composé des  $y_t$  et de chaque  $x_{it}$  ( $x_{it}$  allant de  $x_{1t}$  à  $x_{mt}$ ) est indépendant et identiquement distribué suivant la loi normale multivariée. Alors si  $\mathbf{x}_t^*$  représente un vecteur composé de *n'importe quel* sous-ensemble d'éléments allant de  $x_{1t}$  à  $x_{mt}$ , on peut encore écrire

$$y_t = \beta_0^* + \mathbf{x}_t^* \beta^* + u_t, \quad u_t \sim \text{NID}(0, \sigma_*^2), \quad (2.08)$$

où la notation “ $u_t \sim \text{NID}(0, \sigma_*^2)$ ” est un moyen simple de dire que les  $u_t$  sont **normalement et indépendamment distribués**, ou **n.i.d.**, avec une espérance nulle et une variance égale à  $\sigma_*^2$ . Ceci est vrai pour tout sous-ensemble composé de  $x_{it}$  car toute combinaison linéaire de variables suivant la loi normale multivariée, est elle-même normalement distribuée. Ainsi l'aléa  $u_t$  défini de manière implicite dans (2.08) sera normalement et indépendamment distribué et sans considération des  $x_{it}$  que l'on introduit dans  $\mathbf{x}_t^*$ , et l'on peut toujours choisir  $\beta_0^*$  convenablement de façon à rendre son espérance nulle. Ceci est vrai même si  $\mathbf{x}_t^*$  est un vecteur nul, puisque (2.08) ne fait que traduire l'idée selon laquelle  $y_t$  est égale à son espérance, plus une variable aléatoire  $u_t$  qui est n.i.d. avec une espérance nulle, et  $y_t$  est elle-même normalement distribuée. Pour plus de détails sur ces considérations et sur d'autres cas particuliers, et pour un traitement plus approfondi sur l'interprétation des modèles de régression, consulter Spanos (1986).

Un **modèle** tel que (2.01) devrait être distingué d'un **processus générateur de données**, ou **DGP**, tel que

$$y_t = x_t(\beta_0) + u_t, \quad u_t \sim \text{NID}(0, \sigma_0^2), \quad t = 1, \dots, n. \quad (2.09)$$

Un modèle de régression tel que (2.01) spécifie que l'espérance de  $y_t$  conditionnée par un ensemble *défini* de variables  $\mathbf{Z}_t$  est une fonction *donnée* de  $\mathbf{Z}_t$  et des paramètres (généralement inconnus)  $\beta$ , et que les  $y_t$  sont mutuellement indépendants et ont la même variance autour de leur espérance conditionnelle. D'autre part, un DGP est une caractérisation *complète* des propriétés statistiques de la variable dépendante. Si le DGP est connu, alors aussi bien les valeurs de tous les paramètres que les distributions de toutes les quantités aléatoires doivent être précisées.

Ainsi émergent deux différences fondamentales entre le modèle (2.01) et le DGP (2.09). Le premier implique un vecteur *inconnu* de coefficients  $\beta$ , alors

que l'autre fait référence à un vecteur de coefficients bien défini  $\beta_0$ , qui serait connu si l'on connaissait le DGP. Les aléas  $u_t$  du modèle sont simplement définis comme indépendants et identiquement distribués, avec une espérance nulle et une variance inconnue égale à  $\sigma^2$ , alors que les aléas du DGP sont *normalement* et indépendamment distribués avec une variance connue  $\sigma_0^2$ , qui nous permet de générer une série de  $u_t$  si nous le désirons. Bien évidemment, nous aurions également pu préciser un DGP avec des erreurs qui suivent une distribution autre que la normale; ce qui importe réellement, c'est que la distribution soit spécifiée complètement. D'autre part, nous pouvons être intéressés par ce qui se passe avec la famille entière des DGP, et dans de tels cas une spécification totale n'est pas appropriée.

Un modèle peut ainsi être imaginé comme un **ensemble de DGP**. Lors du processus d'estimation du modèle, ce que nous essayons d'obtenir, c'est une caractérisation estimée du DGP qui a réellement généré les données; dans le cas du modèle de régression non linéaire (2.01) la caractérisation désirée consiste en un ensemble de **paramètres estimés**, c'est-à-dire, des estimations des paramètres inconnus  $\beta$  de la fonction de régression, ainsi qu'une estimation de la **variance des erreurs**,  $\sigma^2$ . Mais puisque dans une régression non linéaire seules l'espérance et la variance des erreurs sont précisées, la caractérisation du DGP obtenue par l'estimation du modèle est *partielle* ou *incomplète*. Plus tard, dans le Chapitre 8, nous discuterons d'une autre méthode d'estimation, celle du maximum de vraisemblance, qui offre une caractérisation complète du DGP après estimation. Ainsi, on peut dire que cette méthode produit un *unique* DGP estimé, alors que toute méthode adoptée pour estimer un modèle de régression non linéaire produit un *ensemble* de DGP, qui satisfont tous la caractérisation estimée.

Cet ensemble de DGP, ou l'unique DGP estimé lorsque ce sera le cas, appartient évidemment à l'ensemble des DGP défini par le modèle. L'estimation statistique peut donc être considérée comme une procédure avec laquelle on sélectionne un sous-ensemble de DGP à partir d'un ensemble donné de DGP. Cette sélection est bien sûr une procédure *aléatoire*, puisqu'un seul DGP appartenant au modèle peut générer des ensembles différents d'observations aléatoires qui entraînent des caractérisations aléatoires estimées différentes. Il est ensuite possible de dissenter sur la *probabilité*, pour un DGP donné, que la caractérisation soit *proche*, dans un certain sens, du DGP lui-même. On peut alors classer ces différentes procédures d'estimation selon ces probabilités, et nous préférons généralement des procédures d'estimation **efficaces**, c'est-à-dire celles pour lesquelles la probabilité que le sous-ensemble sélectionné soit proche du DGP est la plus forte, toujours sous l'hypothèse que le DGP appartient réellement au modèle.

Il nous est impossible de dire quoi que ce soit d'intéressant à propos des propriétés *statistiques* des estimateurs et des statistiques de test sans préciser *à la fois* le modèle *et* le processus qui a généré les données. En pratique bien sûr, nous ne connaissons presque jamais le DGP, sauf si nous procédons à



une expérience Monte Carlo au cours de laquelle nous avons le privilège de générer nous-mêmes les données (consulter le Chapitre 21). Ainsi, lorsque nous estimons des modèles, et à moins d'être extrêmement chanceux, nous ne pouvons pas prétendre raisonnablement que le processus qui a réellement généré les données est un cas particulier du modèle que nous avons estimé, tel que (2.09) l'est de (2.01). Dans le cours que nous développons dans cet ouvrage, nous supposerons néanmoins fréquemment que c'est en fait le cas car il devient alors facile d'établir des résultats définitifs. Mais nous aurons également l'occasion de traiter explicitement des situations où le DGP *n'est pas* un cas particulier du modèle que l'on estime.

La structure additive du modèle de régression non linéaire permet de discuter des deux parties qui composent le modèle séparément. Nous abordons tout d'abord les fonctions de régression, qui déterminent l'espérance conditionnelle de  $y_t$ , et ensuite nous aborderons les aléas qui déterminent tous les moments conditionnels d'ordre supérieur. Il est fondamental de se souvenir que chaque fois que l'on estime un modèle comme (2.01), on fait, implicitement ou explicitement, des hypothèses sur  $x_t(\beta)$  et  $u_t$ , qui sont généralement assez fortes. Puisqu'il est impossible de faire usage des techniques standards pour obtenir des inférences valides si ces hypothèses sont fausses, il est crucial de bien les maîtriser et bien sûr, de les tester contre les valeurs calculées à partir des données.

## 2.5 FONCTIONS DE RÉGRESSION LINÉAIRES ET NON LINÉAIRES

La fonction de régression générale  $x_t(\beta)$  peut être précisée par un grand nombre de moyens. Il peut être très utile de considérer un certain nombre de cas particuliers de façon à avoir une idée de la variété des fonctions de régression spécifiques qui sont le plus souvent utilisées dans la pratique.

La fonction de régression la plus simple est

$$x_t(\beta) = \beta_1 \iota_t = \beta_1, \quad (2.10)$$

où  $\iota_t$  est l'élément  $t$  d'un vecteur dont les  $n$  composantes sont égales à l'unité. Dans ce cas, le modèle (2.01) indique que l'espérance conditionnelle de  $y_t$  est tout simplement une constante. Bien que ce soit un exemple simpliste de fonction de régression, puisque  $x_t(\beta)$  est identique quel que soit  $t$ , il s'agit néanmoins d'un bon exemple pour débiter, et que l'on doit garder à l'esprit. Toutes les fonctions de régression sont tout simplement des versions de (2.10) plus élaborées. Et toute fonction de régression qui ne s'ajuste pas aux données au moins aussi bien que (2.10) devrait être considérée comme une bien mauvaise fonction de régression.

La fonction qui est ensuite la plus simple est la **fonction de régression linéaire simple**

$$x_t(\beta) = \beta_1 + \beta_2 z_t, \quad (2.11)$$

où  $z_t$  est l'unique variable indépendante. En réalité, un modèle encore plus simple consisterait à ne garder que la variable indépendante et à rejeter le terme constant. Cependant, dans la majorité des problèmes appliqués, cela n'a pas de sens d'omettre la constante. De nombreuses fonctions de régression linéaires sont utilisées en tant qu'approximations des fonctions inconnues d'espérance conditionnelle, et de telles approximations seront rarement précises si elles sont contraintes de passer par l'origine. L'équation (2.11) possède deux paramètres, une **ordonnée à l'origine**  $\beta_1$  et une **pente**  $\beta_2$ . Cette fonction est linéaire en ses deux variables ( $\iota_t$  et  $z_t$ , ou tout simplement  $z_t$  si l'on décide de ne pas considérer  $\iota_t$  comme une variable) et en ses paramètres ( $\beta_1$  et  $\beta_2$ ). Bien que ce modèle soit trop simple, il possède certains avantages. Parce qu'il est très facile de grapher  $y_t$  contre  $z_t$ , on peut utiliser ce graphe pour visualiser la fonction de régression, la façon dont le modèle s'ajuste, et si la relation linéaire décrit correctement les données. Mais lorsqu'un modèle intègre plus d'une variable indépendante, visualiser les données de cette façon devient plus problématique, et donc moins habituel.

Une généralisation évidente de (2.11) est la **fonction de régression linéaire multiple**

$$x_t(\boldsymbol{\beta}) = \beta_1 z_{t1} + \beta_2 z_{t2} + \beta_3 z_{t3} + \cdots + \beta_k z_{tk}, \quad (2.12)$$

où les  $z_{ti}$  ( $z_{ti}$  allant de  $z_{t1}$  à  $z_{tk}$ ) sont les variables indépendantes, et  $z_{t1}$  peut être un terme constant. Il aurait été possible de formuler cette fonction de régression de façon plus ramassée

$$x_t(\boldsymbol{\beta}) = \mathbf{Z}_t \boldsymbol{\beta},$$

où  $\mathbf{Z}_t$  représente un vecteur de dimension  $1 \times k$ , et  $\boldsymbol{\beta}$  désigne un vecteur de dimension  $k \times 1$ . Notons que (2.12) repose sur une hypothèse extrêmement forte, c'est-à-dire celle que l'effet sur  $y_t$  d'une modification d'une des variables indépendantes est indépendant des valeurs de toutes les autres variables indépendantes. Lorsque cette hypothèse est fautive, les modèles de régression linéaire multiple peuvent sérieusement induire une erreur.

Puis vient tout un éventail de fonctions de régression ressemblant à

$$x_t(\boldsymbol{\beta}) = \beta_1 z_{t1} + \beta_2 z_{t2} + \beta_3 z_{t2}^2 + \beta_4 z_{t1} z_{t2},$$

qui est linéaire en ses paramètres mais qui fait appel à des variables indépendantes d'une manière non linéaire. Les modèles qui impliquent cette famille de fonctions de régression peuvent être manipulés comme n'importe quel autre modèle de régression linéaire, tout simplement en définissant de nouveaux régresseurs de façon appropriée. Ici, par exemple, on pourrait définir  $z_{t3}$  comme  $z_{t2}^2$  et  $z_{t4}$  comme  $z_{t1} z_{t2}$ . En faisant usage de ce genre de fonction on évite de subir les effets qui s'additionnent, comme l'implique (2.12), mais cela nécessiterait sans doute d'estimer plus de paramètres qu'il ne serait utile en pratique avec de nombreux ensembles de données. A cause de cela, et à moins

qu'il n'existe des raisons théoriques de s'attendre à ce que des puissances ou des produits de variables indépendantes n'apparaissent dans la fonction de régression, la plupart des économètres essaieront d'ignorer ce genre de spécification en pratique.

Une fonction de régression qui permet à toutes les variables indépendantes d'interagir sans recourir à l'estimation de paramètres supplémentaires est la fonction multiplicative

$$x_t(\beta) = e^{\beta_1 z_{t2}^{\beta_2} z_{t3}^{\beta_3}}. \quad (2.13)$$

Remarquons que cette fonction peut être évaluée uniquement lorsque  $z_{t2}$  et  $z_{t3}$  sont positifs pour tout  $t$ . C'est la première véritable fonction de régression non linéaire que nous rencontrons, puisqu'il est clair qu'elle n'est linéaire ni en ses paramètres ni en ses variables. Cependant, un modèle non linéaire tel que

$$y_t = e^{\beta_1 z_{t2}^{\beta_2} z_{t3}^{\beta_3}} + u_t \quad (2.14)$$

est très rarement estimé dans la pratique. La raison en est que l'hypothèse d'aléas additifs et identiquement distribués est autant encombrante que peu réaliste. Elle est peu réaliste car les  $z_{ti}$  sont multiplicatifs, ce qui implique que leurs effets dépendent des niveaux que prennent toutes les valeurs des autres variables, alors que les aléas sont additifs, ce qui rend leur effet indépendant des niveaux des autres variables explicatives. Elle est encombrante car (2.14) doit être estimée par moindres carrés non linéaires plutôt que par moindres carrés linéaires.

Il est facile de modifier (2.14) de façon à donner aux aléas une structure multiplicative. Le modèle le plus évident que l'on peut alors formuler est

$$y_t = (e^{\beta_1 z_{t2}^{\beta_2} z_{t3}^{\beta_3}})(1 + v_t) \equiv e^{\beta_1 z_{t2}^{\beta_2} z_{t3}^{\beta_3}} + u_t, \quad (2.15)$$

où les perturbations  $1 + v_t$ , qui sont des quantités sans unité de mesure, sont multiplicatives. Bien que les erreurs sous-jacentes  $v_t$  soient i.i.d., les erreurs additives  $u_t$  sont maintenant proportionnelles à la fonction de régression. Si le modèle s'ajuste relativement bien, les  $v_t$  devraient être assez faibles (disons inférieures à environ 0.05). Maintenant, souvenons-nous que  $e^w \cong 1 + w$  pour des valeurs de  $w$  proches de zéro. Par conséquent, pour des modèles qui s'ajustent relativement bien, (2.15) sera très similaire au modèle

$$y_t = e^{\beta_1 z_{t2}^{\beta_2} z_{t3}^{\beta_3}} e^{v_t}. \quad (2.16)$$

Supposons désormais que l'on passe en logarithme, de chaque côté de l'égalité. Le résultat est

$$\log(y_t) = \beta_1 + \beta_2 \log(z_{t2}) + \beta_3 \log(z_{t3}) + v_t, \quad (2.17)$$

qui est un modèle de régression linéaire. Il est évident que ce modèle, qui est linéaire dans tous les paramètres et dans les logarithmes de toutes les

variables, sera plus facile à estimer que le modèle non linéaire (2.14). Les arguments que l'on a développés plus tôt suggèrent que c'est, en tout cas, plus plausible. Ainsi, il ne devrait pas être surprenant d'apprendre que les modèles de régression log-linéaire, comme (2.17), sont très fréquemment estimés en pratique, alors que les modèles multiplicatifs avec des aléas additifs comme (2.14) ne le sont que très rarement.

Un modèle purement multiplicatif comme (2.16) peut être rendu linéaire en passant en logarithme. Toutefois, un modèle qui mélange les deux structures, multiplicative et additive, ne peut pas être transformé en un modèle linéaire. Ainsi, peu importe la manière dont sont précisés les aléas; des modèles qui intègrent des fonctions de régression du type

$$x_t(\beta) = \beta_1 + \beta_2 z_{t2}^{\beta_3} + \beta_4 z_{t3} \quad \text{et} \quad (2.18)$$

$$x_t(\beta) = \beta_1 + \beta_2 z_{t2}^{\beta_3} z_{t3}^{\beta_4} \quad (2.19)$$

doivent nécessairement être estimés à l'aide des méthodes non linéaires. Comme on devrait s'y attendre, de tels modèles ne sont pas estimés aussi fréquemment que les modèles linéaires ou log-linéaires, d'une part parce que la paresse nous y pousse sans doute, et d'autre part car il n'y a souvent pas de raison, ni théorique ni empirique, qui nous permettent de choisir ce type de spécification plutôt que les modèles conventionnels. En fait, les fonctions de régression comme (2.18) et (2.19) sont d'une difficulté de traitement notoire, car il est complexe d'estimer conjointement tous les paramètres avec n'importe quel degré de précision. Souvenons-nous de la discussion à propos du fait que les modèles fondés sur la fonction de régression (2.06), qui est très similaire à celles-ci, sont le plus souvent insuffisamment identifiés.

L'ultime exemple d'une fonction de régression non linéaire que nous allons aborder est très différent par rapport à (2.18). Considérons la fonction de régression

$$x_t(\beta) = \beta_1 + \beta_2(z_{t2} - \beta_3 z_{t3}) + \beta_4(z_{t4} - \beta_3 z_{t5}). \quad (2.20)$$

Cette fonction est linéaire en ses variables indépendantes  $z_t$  et  $z_{t2}, z_{t3}, z_{t4}$  et  $z_{t5}$ , mais elle est non linéaire en ses paramètres  $\beta_i$  (allant de  $\beta_1$  à  $\beta_4$ ). Mais il s'agit en réalité d'une fonction de régression linéaire avec une seule **contrainte non linéaire** sur les coefficients. Pour apercevoir ceci, examinons la fonction de régression linéaire non contrainte

$$x_t(\beta) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \gamma_5 z_{t5}.$$

Si l'on impose la contrainte non linéaire

$$\frac{\gamma_3}{\gamma_5} = \frac{\gamma_2}{\gamma_4}, \quad (2.21)$$

et si l'on reparamétrise ensuite de façon à ce que

$$\beta_1 = \gamma_1, \quad \beta_2 = \gamma_2, \quad \beta_3 = -\frac{\gamma_5}{\gamma_4}, \quad \text{et} \quad \beta_4 = \gamma_4,$$

on obtient (2.20). Remarquons qu'il y a plusieurs manières équivalentes d'écrire la contrainte (2.21), dont

$$\gamma_3 = \frac{\gamma_2\gamma_5}{\gamma_4}, \quad \gamma_2 = \frac{\gamma_3\gamma_4}{\gamma_5}, \quad \text{et} \quad \frac{\gamma_2}{\gamma_3} = \frac{\gamma_4}{\gamma_5}.$$

Il s'agit d'un caractère typique des contraintes non linéaires que de pouvoir être formulées de plusieurs façons différentes mais équivalentes, et par conséquent, la fonction de régression peut être paramétrisée de différentes façons.

On retrouve très fréquemment des fonctions de régression comme (2.20) en économétrie. Elles apparaissent, par exemple, dans certains modèles avec anticipations rationnelles — consulter Hoffman et Schmidt (1981) ou Gregory et Veall (1985, 1987) — et dans les modèles avec corrélation en série (voir Chapitre 10). De tels modèles ne sont pas particulièrement difficiles à estimer en général, pourvu que les contraintes soient plus ou moins exactes.

## 2.6 TERMES D'ALÉA

Il existe deux éléments que l'on doit préciser lorsque l'on spécifie un modèle de régression: la fonction de régression  $x_t(\beta)$  et au moins quelques propriétés des aléas  $u_t$ . Nous avons déjà eu l'occasion de constater à quel point ces dernières étaient importantes. En rajoutant les erreurs à variance constante à la fonction de régression à structure multiplicative (2.13), nous avons obtenu un modèle de régression véritablement non linéaire. Mais lorsque nous avons appliqué des erreurs qui étaient proportionnelles à la fonction de régression, comme dans (2.15), et fait usage de l'approximation  $e^w \cong 1 + w$ , qui est une approximation satisfaisante pour des petites valeurs de  $w$ , nous avons obtenu un modèle de régression log-linéaire. Il devrait donc être clair à partir de cet exemple, que la manière dont sont précisés les aléas aura un effet considérable sur le modèle qui est réellement estimé.

Dans (2.01), nous avons défini les aléas comme indépendants, tous d'espérance nulle et de variance égale à  $\sigma^2$ , mais nous n'avons pas précisé leur distribution. Même ces hypothèses sont quelquefois trop fortes. Elles excluent toutes les sortes de dépendance à travers les observations, et toutes les sortes de variation dans le temps ou avec les valeurs de n'importe quelle variable indépendante. Elles excluent également des distributions où les queues sont tellement épaisses que les aléas n'ont pas une variance finie. Une telle distribution est la distribution de Cauchy. Une variable aléatoire qui suit une distribution de Cauchy ne possède pas seulement une variance non finie, mais aussi une espérance non finie. Consulter le Chapitre 4 et l'Annexe B.

Il existe plusieurs acceptions du terme *indépendance* dans la littérature consacrée à la statistique et à l'économétrie. Deux variables aléatoires  $z_1$  et  $z_2$  sont dites **aléatoirement indépendantes** si leur fonction de répartition

jointe  $F(z_1, z_2)$  est égale au produit de leurs deux fonctions de répartition marginale respectives  $F(z_1, \infty)$  et  $F(\infty, z_2)$ . On appelle quelquefois cela l'**indépendance en probabilité**, mais nous ferons usage du premier terme, plus moderne. Certains auteurs écrivent que deux variables aléatoires  $z_1$  et  $z_2$  sont **linéairement indépendantes** si  $E(z_1 z_2) = E(z_1)E(z_2)$ , une condition moins forte, qui découle de l'indépendance stochastique, mais qui ne l'entraîne pas. Cette terminologie est assez malvenue car le sens "linéairement indépendant" ne s'accorde pas avec le sens habituel que l'on utilise en algèbre linéaire. Au contraire, dans cette situation, on pourrait au plus dire que  $z_1$  et  $z_2$  sont **non corrélées**, et possèdent une covariance nulle. Si  $z_1$ , ou  $z_2$ , est d'espérance nulle et est non corrélée avec  $z_2$  (respectivement  $z_1$ ), alors  $E(z_1 z_2) = 0$ . Il existe un sens selon lequel  $z_1$  et  $z_2$  sont **orthogonaux** dans cette situation, et nous utiliserons quelquefois cette terminologie.

Lorsque nous disons que les  $u_t$  sont i.i.d., nous signifions par le premier "i" que les  $u_t$  sont aléatoirement indépendants. Cela implique que  $E(u_t u_s) = 0$  pour tout  $t \neq s$ , mais également que  $E(h_1(u_t)h_2(u_s)) = 0$  pour toutes les fonctions (mesurables)  $h_1(\cdot)$  et  $h_2(\cdot)$ . Les aléas qui sont indépendants et qui possèdent les mêmes espérances et variances sont quelquefois appelés **bruits blancs**. Cette terminologie que l'on emprunte à la littérature scientifique, se réfère au fait que, tout comme la lumière blanche est constituée de quantités égales de rayonnements de toutes les parties du spectre visible, les erreurs bruits blancs contiennent des quantités égales d'aléas de toutes fréquences. De nombreuses définitions différentes des bruits blancs sont en usage en économétrie et dans d'autres disciplines, et quelquefois, le terme est employé dans un sens qui n'est pas strictement conforme à sa signification.

Remarquons l'importante distinction qu'il faut établir entre les aléas et les **résidus**. *Toute* régression linéaire ou non linéaire génère un vecteur de résidus, que cela ait un sens ou pas. Les résidus auront des propriétés qui résultent de la façon dont on les a obtenus, sans se préoccuper de la manière dont les données ont été générées. Par exemple, les résidus OLS seront toujours orthogonaux à tous les régresseurs, et les résidus NLS seront toujours orthogonaux à la matrice  $\hat{X}$ . D'un autre côté, les aléas ne sont pas observables (mais on peut les estimer) et l'on doit formuler quelques hypothèses qui feront partie de la définition du modèle. Il nous arrivera bien sûr de tester ces hypothèses, et de le faire à l'aide de statistiques de tests dépendant des résidus que l'on calculera.

Une grande partie de la littérature concernant la spécification et les tests des modèles de régression est consacrée aux tests de transgression des hypothèses d'erreurs i.i.d.. Lorsque de telles hypothèses ne sont pas bien vérifiées, il est encore possible de modifier le modèle avec des erreurs qui ne sont pas i.i.d. en un modèle où les erreurs transformées le sont. Il se peut que l'hypothèse d'indépendance, ou que l'hypothèse d'espérances et de variances identiques, ou les deux simultanément, ne soient pas vérifiées. L'hypothèse d'indépendance est quelquefois mise en défaut lorsque l'on travaille sur des

**données chronologiques:** les aléas successifs  $u_t$  peuvent être corrélés entre eux, faisant apparaître plus distinctement le phénomène de **corrélacion en série**. L'hypothèse de distributions identiques est souvent mise à mal lorsque l'on travaille avec des **données en coupe transversale:** des  $u_t$  différents peuvent sembler provenir de la même famille de distribution mais ont des variances différentes, et mettent en perspective le phénomène d'**hétéroscédasticité**. Le terme opposé hétéroscédasticité est incidemment homoscedasticité. Si les aléas possèdent une variance commune, on dit qu'ils sont **homoscedastiques**; lorsque ce n'est pas le cas on dit qu'ils sont **hétéroscédastiques**. Bien sûr, la corrélation des aléas à travers les observations n'est en rien une caractéristique exclusive des données chronologiques, et l'hétéroscédasticité n'est en rien une caractéristique exclusive des données en coupe transversale. Ces deux phénomènes peuvent survenir avec tous les types d'ensembles de données, mais malgré tout, on associe nécessairement la corrélation *en série* avec les données chronologiques, et l'hétéroscédasticité est particulièrement fréquente avec les données en coupe transversale.

Nous traiterons plus en détail la corrélation en série et l'hétéroscédasticité dans les chapitres qui leur sont consacrés (tout particulièrement, les Chapitres 9, 10, 11 et 16). Pour l'instant, et à titre d'illustration, considérons une forme simple d'hétéroscédasticité:

$$u_t = w_t v_t, \quad v_t \sim \text{IID}(0, \sigma_v^2),$$

où  $w_t$  est une variable indépendante qui est toujours non nulle. Cette spécification implique que  $u_t$  possède une espérance nulle et une variance égale à  $\sigma_v^2 w_t^2$ . Supposons désormais que la fonction de régression sur laquelle on applique les erreurs  $u_t$  soit

$$x_t(\beta) = \beta_1 + \beta_2 z_t + \beta_3 w_t.$$

Bien évidemment, on peut obtenir un modèle avec des erreurs i.i.d. en divisant la variable dépendante et toutes les variables indépendantes, la constante comprise par  $w_t$ . Ce modèle modifié est

$$\frac{y_t}{w_t} = \beta_1 \frac{1}{w_t} + \beta_2 \frac{z_t}{w_t} + \beta_3 + v_t. \quad (2.22)$$

Notons que les régresseurs sont désormais  $1/w_t$ ,  $z_t/w_t$ , et une constante, mais le coefficient de la constante est maintenant celui de  $w_t$  dans le modèle originel, alors que le coefficient  $1/w_t$  est la constante du modèle de départ. Ainsi il est très facile d'éliminer l'hétéroscédasticité dans un cas pareil, mais il faut être prudent en interprétant les coefficients du modèle transformé.

Au Chapitre 8, nous discuterons d'une hypothèse relativement forte que l'on fait en économétrie, c'est-à-dire

$$u_t \sim \text{NID}(0, \sigma^2), \quad t = 1, \dots, n,$$

qui précise que les  $u_t$  sont **normalement et indépendamment distribués** avec une espérance nulle et une variance égale à  $\sigma^2$ . Ainsi chaque  $u_t$  est supposé obéir à la distribution normale dont la fonction de densité de probabilité est

$$f(u_t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{u_t^2}{2\sigma^2}\right).$$

La densité jointe du vecteur à  $n$  composantes  $\mathbf{u}$  (dont l'élément type est  $u_t$ ) est supposée être par conséquent

$$f(\mathbf{u}) = \prod_{t=1}^n f(u_t) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^n u_t^2\right).$$

Il existe trois raisons principales pour supposer la normalité. La première d'entre elles est que grâce à leur facilité de calcul et à leurs propriétés familières, on désire souvent faire usage des moindres carrés pour estimer des modèles de régression, et la justification de cet usage est plus solide lorsque les erreurs sont normalement distribuées que lorsque ce n'est pas le cas. Comme nous le verrons au cours du Chapitre 8, les moindres carrés appliqués à un modèle de régression disposent de propriétés asymptotiques excellentes lorsque les erreurs sont normales, mais lorsque ces erreurs suivent une autre distribution quelconque connue, leurs propriétés ne sont plus aussi bonnes. La deuxième raison est que lorsque l'on suppose la normalité, on peut obtenir le plus souvent des résultats plus solides que lorsque l'on suppose simplement que les erreurs sont supposées être i.i.d.. En particulier, pour les modèles de régression linéaire avec régresseurs fixés et erreurs normales, nous pouvons obtenir des résultats exacts avec des échantillons finis (consulter le Chapitre 3); de tels résultats ne sont même pas disponibles pour des modèles linéaires quand les erreurs sont simplement supposées être i.i.d.. La troisième raison est que lorsque l'on quitte le domaine des modèles de régression pour essayer de traiter des modèles non linéaires plus généraux, il devient souvent nécessaire de faire des hypothèses sur la distribution, et la distribution normale est bien souvent la plus pratique à utiliser.

Aucune de ces raisons pratiques de supposer que les aléas sont normalement distribués n'offre une quelconque *justification* pour formuler une telle hypothèse. L'argument usuel est que les aléas représentent les effets combinés de nombreuses variables que l'on a oubliées, et les nombreuses erreurs de mesure. Les **Théorèmes de la Limite Centrale** (que nous verrons au Chapitre 4) nous affirment que, très grossièrement, lorsque l'on établit la moyenne d'un grand nombre de variables aléatoires, la moyenne obtenue est approximativement normalement distribuée, en rapport plus ou moins fidèle avec les distributions des variables aléatoires originelles. L'argument habituel est que l'hypothèse de normalité a du sens parce que nous pouvons penser que les aléas dans les modèles de régression sont ainsi en moyenne.



Il y a au moins deux problèmes avec ce genre d'argument. Premièrement, comme nous le verrons au Chapitre 4, les théorèmes de la limite centrale nécessitent des hypothèses relativement fortes. Ils s'appliquent à des situations où l'on fait la moyenne de plusieurs variables aléatoires, dont aucune n'est "grande" par rapport à toutes les autres. Il est aisé de penser à des variables économiques qui peuvent être omises dans les modèles de régression, et qui constituent donc une partie des aléas, mais qui seraient peut-être relativement importantes par rapport à ces aléas. Dans le cas de modèles chronologiques, des grèves, des élections ou d'autres événements politiques, et des tempêtes ou d'autres conditions climatiques extrêmes sont quelques exemples qui nous viennent à l'esprit. Il n'existe sans doute aucune raison *a priori* de s'attendre à ce que les effets de tels événements ne soient responsables que d'une petite partie de l'erreur globale pour toute observation donnée. Dans le cas des modèles à coupe transversale, l'argument de normalité est probablement moins pesant. Lorsque nous disposons d'un échantillon important d'individus ou d'entreprises, nous devons constater que quelques observations comprises dans l'échantillon ne doivent pas s'y trouver en réalité. Considérons, par exemple, l'effet sur un modèle en coupe transversale de la demande de viande d'un petit nombre d'individus végétariens! Inévitablement, les erreurs associées à ces observations particulières seront élevées, de sorte qu'il est peu probable que la distribution des aléas pour le modèle tout entier soit normale.

Le second problème avec l'argument du théorème de la limite centrale est que beaucoup de théorèmes de la limite centrale ne s'appliquent pas lorsque le nombre de variables aléatoires dont on fait la moyenne est lui-même aléatoire. Mais puisque nous ne savons pas quelles variables ont été omises et rejetées dans les aléas, nous n'avons aucune raison d'imaginer que leur nombre est le même d'observation en observation! Alors on ne peut pas toujours invoquer légitimement un théorème de la limite centrale.

Ces arguments ne doivent pas suggérer qu'il est idiot de supposer la normalité. Mais que nous ayons supposé ou pas la normalité ne nous empêche pas de voir si oui ou non les aléas sont en réalité approximativement normaux. Si ils ne sont pas approximativement normaux, alors la sagesse nous conseille de remettre en question l'usage des moindres carrés. Il existe, bien sûr, un nombre infini de distributions non normales, et donc un nombre infini de types de non normalité à examiner. Cependant, la grande majorité des tests de non normalité mettent l'accent sur deux propriétés de la distribution normale. Si  $\varepsilon \sim N(\mu, \sigma^2)$ , alors

$$E((\varepsilon - \mu)^3) = 0 \quad \text{et} \quad (2.23)$$

$$E((\varepsilon - \mu)^4) = 3\sigma^4. \quad (2.24)$$

L'expression (2.23) nous renseigne que pour la distribution normale, le troisième **moment centré** (c'est-à-dire, le moment centré autour de l'espérance) est nul. Ce moment est fréquemment utilisé pour mesurer l'**asymétrie**. Positif, il indique que la distribution est *biaisée* à droite; négatif, il indique que

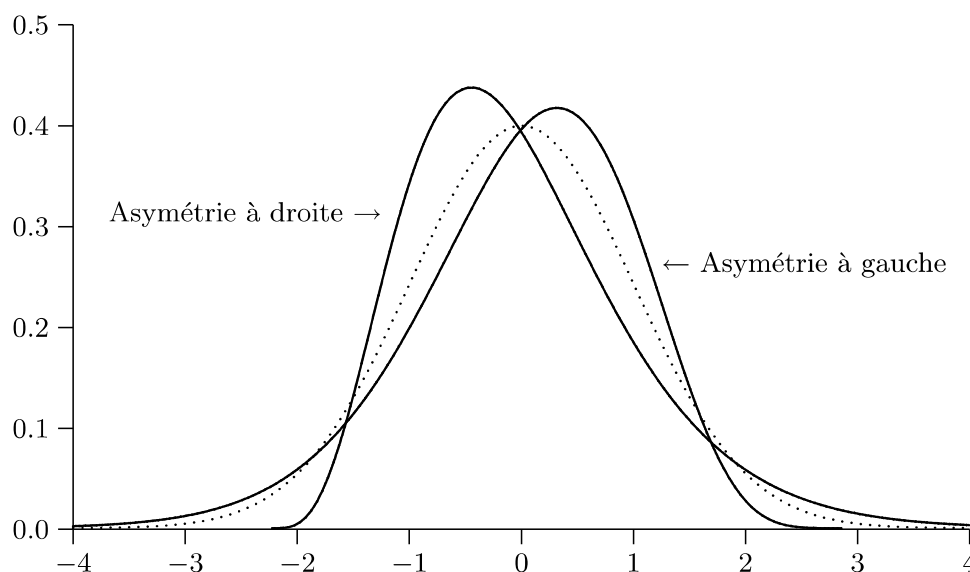
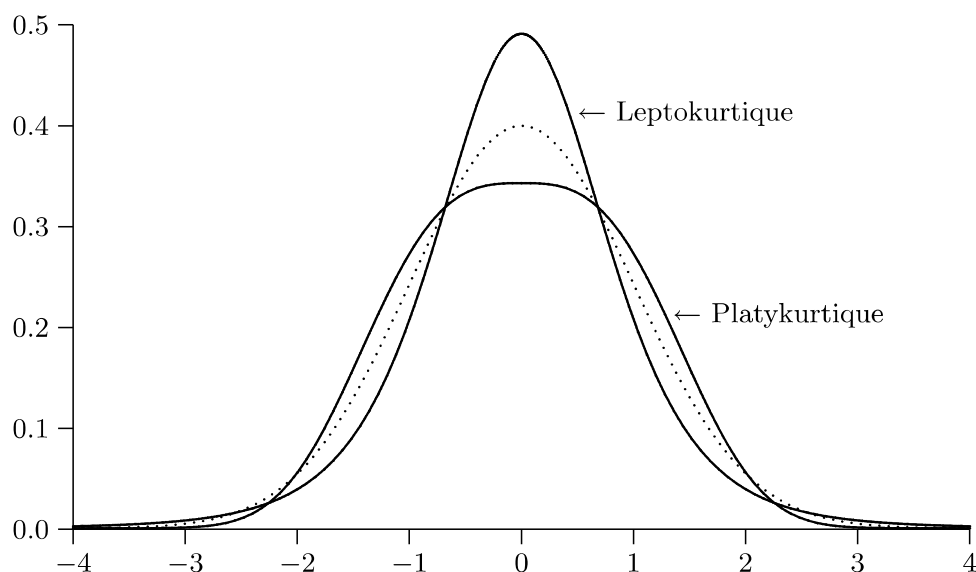


Figure 2.8 Distributions asymétriques

la distribution est *biaisée* à gauche. La Figure 2.8 illustre deux distributions asymétriques, et, pour comparer une distribution symétrique. Les tests d'asymétrie sont relativement aisés; ils seront traités dans le chapitre 16.

L'expression (2.24) nous indique que le quatrième moment centré d'une variable aléatoire normale est égale à trois fois le carré de sa variance. Une variable aléatoire dont le quatrième moment est plus élevé que trois fois le carré de son deuxième moment possède des queues de distribution plus épaisses qu'une variable aléatoire qui suit une distribution normale. On dit quelquefois qu'il fait état de l'**excès de kurtosis** ou que la distribution est **leptokurtique**. Au contraire, lorsqu'une variable aléatoire a un quatrième moment inférieur à trois fois le carré de son second moment, elle possède des queues de distribution plus fines qu'une variable aléatoire distribuée normalement. De telles variables aléatoires sont dites **platykurtiques**. De façon similaire, on dit souvent des variables aléatoires qui suivent la distribution normale qu'elles sont **mésokurtiques**. Les lecteurs qui ont quelques notions de Grec pourraient penser que ces définitions sont erronées, puisque *lepto* signifie *fin* et *platy* signifie *épais*. Comme l'expliquent Kendall et Stuart (1977, p. 88), ces termes étaient à l'origine destinés à caractériser les parties centrales des distributions et non les queues de distribution; ainsi les distributions leptokurtiques sont ainsi nommées non pas parce qu'elles ont des queues de distributions épaisses mais parce qu'elles ont des parties centrales (relativement) minces, et les distributions platykurtiques sont ainsi nommées non pas à cause de leurs queues de distribution fines parce qu'elles ont des parties centrales (relativement) épaisses. Toutefois, ce sont aux queues de distribution auxquelles se réfèrent les statisticiens contemporains en employant ces termes. La Figure 2.9 illustre



**Figure 2.9** Distributions leptokurtique et platykurtique

des distributions leptokurtiques et platykurtiques. A titre de comparaison, la distribution normale standard a également été représentée (en pointillé).

Les queues de distribution fines ne représentent pas vraiment un problème (et ne sont pas non plus très fréquentes), mais les queues de distribution épaisses peuvent causer de graves difficultés pour l'estimation et l'inférence. Si les aléas suivent une distribution dont les queues de distribution sont plus épaisses que celles de la distribution normale, alors des erreurs importantes inhabituelles surviendront relativement souvent. La procédure des moindres carrés donne un grand poids à ces erreurs importantes, et peut donc entraîner des estimations des paramètres inefficaces.

Il est assez facile de tester l'excès de kurtosis; voir Chapitre 16. Cependant, ce qu'il faut faire si l'on trouve un excès de kurtosis substantiel, n'est pas clairement établi. L'hétéroscédasticité peut conduire à l'*apparence* de kurtosis, comme le ferait une fonction de régression incorrectement spécifiée, de sorte qu'il serait souhaitable d'examiner la spécification du modèle. Si l'on est confiant dans la spécification de la fonction de régression et qu'il n'y a pas d'hétéroscédasticité, alors il serait sûrement plus sage de considérer d'autres méthodes que les moindres carrés. Il existe une littérature importante consacrée à ce que les statisticiens appellent des méthodes d'estimations "robustes", qui donnent un poids plus faible aux valeurs détachées que les moindres carrés; consulter Krasker, Kuh, et Welsch(1983) pour une revue de littérature. De manière alternative, on pourrait postuler d'autres distributions que la normale qui possèderait des queues de distribution plus épaisses, puis faire usage de la méthode du maximum de vraisemblance, dont nous discuterons en détail au cours du Chapitre 8 et des chapitres suivants.

## 2.7 CONCLUSION

Ce chapitre nous a donné une introduction non rigoureuse aux modèles de régression non linéaire, mettant l'accent sur des concepts fondamentaux tels que la géométrie de la régression non linéaire. Les ouvrages qui offrent un traitement plus rigoureux sont ceux de Gallant (1987), Bates et Watts (1988), et Seber et Wild (1989). Le prochain chapitre traite de la façon d'opérer des inférences à partir de modèles de régression non linéaire et introduit les idées de base des tests d'hypothèses pour de tels modèles. La prochaine étape devra offrir un traitement des propriétés asymptotiques des moindres carrés non linéaires, et cela sera l'objet des Chapitres 4 et 5. Puis le Chapitre 6 examinera une régression linéaire "artificielle" de Gauss-Newton que l'on associe à tout modèle de régression non linéaire. Cette régression artificielle s'avèrera très utile pour toute une variété d'usages, dont le calcul des estimations NLS et le calcul des statistiques de test.

## TERMES ET CONCEPTS

aléas	modèles de régression: linéaire et non
algorithme de minimisation	linéaire, multivariée et univariée
bruit blanc	moindres carrés non linéaires
colinéarité	moments centrés
corrélation en série	moyenne conditionnelle
distribution normale	multicolinéarité
données chronologiques	asymétrie
données en coupe transversale	processus générateur de données
ensemble d'informations	(DGP); relation avec les modèles
fonction somme des carrés	restrictions non linéaires
hétéroscédasticité	résultats asymptotiques
homoscédasticité	Théorèmes de la Limite Centrale
identification: globale et locale	variables aléatoires indépendantes et
indépendance: stochastique et linéaire	identiquement distribuées (i.i.d.)
kurtosis: leptokurtique, mésokurtique,	variables dépendantes et
platykurtique, excès de kurtosis	indépendantes
minima: locaux et globaux	variance d'erreur
modèle: ensemble de DGP	

# Chapitre 3

## Inférence dans les Modèles de Régression non Linéaire

### 3.1 INTRODUCTION

Supposons que l'on dispose d'un vecteur  $\mathbf{y}$  donné des observations de quelques variables dépendantes, d'un vecteur  $\mathbf{x}(\boldsymbol{\beta})$  de fonctions de régression en général non linéaires, qui devraient dépendre et dépendront en général de variables indépendantes, et des données nécessaires au calcul de  $\mathbf{x}(\boldsymbol{\beta})$ . Puis en supposant que ces données permettent d'identifier tous les éléments du vecteur de paramètres  $\boldsymbol{\beta}$ , et que l'on ait accès à un programme informatique adéquat pour les moindres carrés non linéaires et à suffisamment de temps, il serait possible d'obtenir des estimations NLS  $\hat{\boldsymbol{\beta}}$ . Dans le but d'interpréter ces estimations, il faut généralement faire l'hypothèse très forte selon laquelle  $\mathbf{y}$  est en réalité généré par un DGP de la famille

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (3.01)$$

de sorte que le modèle est supposé “correct”. Sans recourir à cette hypothèse, ou à une de ses variantes moins restrictives, il serait extrêmement difficile de dire quoi que ce soit à propos des propriétés de  $\hat{\boldsymbol{\beta}}$ , bien que cela soit possible dans certains cas particuliers.

Il est clair que  $\hat{\boldsymbol{\beta}}$  doit être un vecteur de variables aléatoires, puisqu'il dépend de  $\mathbf{y}$ , et donc du vecteur d'aléas  $\mathbf{u}$ . Ainsi, si l'on désire opérer des inférences sur  $\boldsymbol{\beta}$ , il nous faut reconnaître que  $\hat{\boldsymbol{\beta}}$  est aléatoire, et quantifier cet aléa. Dans le Chapitre 5, nous démontrerons lorsque la taille de l'échantillon est suffisamment importante qu'il est raisonnable de traiter  $\hat{\boldsymbol{\beta}}$  comme normalement distribué autour de la vraie valeur de  $\boldsymbol{\beta}$ , que nous nommerons  $\boldsymbol{\beta}_0$ . Ainsi, la seule chose dont on a besoin si l'on désire pratiquer des inférences asymptotiques correctes à propos de  $\boldsymbol{\beta}$  est la **matrice de covariance** de  $\hat{\boldsymbol{\beta}}$ , notée  $\mathbf{V}(\hat{\boldsymbol{\beta}})$ . Dans la prochaine section, nous verrons comment la matrice de covariance pourrait être estimée pour les modèles de régression linéaire et non linéaire, et dans la Section 3.3 nous montrerons comment les estimations obtenues peuvent être utilisées pour faire des inférences sur  $\boldsymbol{\beta}$ . Dans la Section 3.4, nous discuterons des idées de base qui sous-tendent tous les types

de tests d'hypothèses. Dans la Section 3.5, nous discuterons des procédures pour les tests d'hypothèses dans les modèles de régression linéaire, et dans la Section 3.6, nous verrons les procédures similaires qui s'appliquent aux tests d'hypothèses dans les modèles de régression non linéaire. Cette dernière section nous offrira l'opportunité de présenter les trois principes fondamentaux sur lesquels la plupart des tests d'hypothèses se basent, c'est-à-dire les principes de Wald, du multiplicateur de Lagrange et du rapport de vraisemblance. Enfin, dans la Section 3.7, nous discuterons des effets de restrictions incorrectes, et nous introduirons la notion des estimateurs issus des tests préliminaires.

### 3.2 ESTIMATION DE LA MATRICE DE COVARIANCE

Dans le cas du modèle de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (3.02)$$

il est bien connu que lorsque le DGP vérifie (3.02) pour des valeurs définies des paramètres de  $\boldsymbol{\beta}_0$  et de  $\sigma_0$ , la matrice de covariance du vecteur des estimations  $\hat{\boldsymbol{\beta}}$  obtenu par OLS, est

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \quad (3.03)$$

La démonstration de ce résultat familier est assez immédiate. La matrice de covariance  $\mathbf{V}(\hat{\boldsymbol{\beta}})$  est définie comme l'espérance du produit extérieur de  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$  avec lui-même, conditionnellement aux variables indépendantes  $\mathbf{X}$ . À partir de cette définition, et en utilisant le fait que  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ , on remplace tout d'abord  $\hat{\boldsymbol{\beta}}$  par un terme qui lui est équivalent sous le DGP, puis on calcule l'espérance conditionnellement à  $\mathbf{X}$ , et on simplifie l'expression algébrique pour obtenir (3.03):

$$\begin{aligned} \mathbf{V}(\hat{\boldsymbol{\beta}}) &\equiv E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \\ &= E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta}_0)((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \boldsymbol{\beta}_0)^\top \\ &= E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}) - \boldsymbol{\beta}_0)((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}) - \boldsymbol{\beta}_0)^\top \\ &= E(\boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} - \boldsymbol{\beta}_0)^\top \\ &= E(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\sigma_0^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

Obtenir un résultat analogue pour le modèle de régression non linéaire (3.01) requiert quelques concepts d'analyse asymptotique que nous n'avons pas

développés jusqu'ici, et une certaine quantité de manipulations mathématiques. Nous repoussons donc cette analyse au Chapitre 5, et nous établissons ici un résultat au plus approximatif.

Pour un modèle non linéaire, on ne peut pas en général obtenir une expression exacte de  $V(\hat{\beta})$  dans le cas d'un échantillon de taille finie. Dans le Chapitre 5, sous l'hypothèse que les données ont été générées par un DGP qui serait un cas particulier de (3.01), nous obtiendrons toutefois un résultat asymptotique qui nous permettra d'établir que

$$V(\hat{\beta}) \cong \sigma_0^2 (\mathbf{X}^\top(\beta_0) \mathbf{X}(\beta_0))^{-1}, \quad (3.04)$$

où  $\cong$  signifie “est approximativement égal à”, et  $\mathbf{X}(\beta_0)$  est la matrice des dérivées partielles de la fonction de régression présentée dans (2.04). La qualité de cette approximation dépendra du modèle et de la taille de l'échantillon. Elle sera en général bien meilleure pour des modèles quasiment linéaires et pour des échantillons de grande taille.

Dans la pratique, nous ne pourrions bien évidemment pas faire usage de (3.04) car  $\sigma_0^2$  et  $\beta_0$  ne sont pas connus; il nous faut les estimer. Le seul moyen raisonnable d'estimer  $\beta_0$  dans ce contexte est de prendre  $\hat{\beta}$ , mais il y a au moins deux façons d'estimer  $\sigma_0^2$ . Il en résulte deux façons d'estimer  $V(\hat{\beta})$ . La première que l'on peut utiliser est

$$\hat{V}(\hat{\beta}) \equiv \hat{\sigma}^2 (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1}, \quad (3.05)$$

où  $\hat{\sigma}^2 \equiv n^{-1} SSR(\hat{\beta})$ , et l'autre est

$$\mathbf{V}_s(\hat{\beta}) \equiv s^2 (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1}, \quad (3.06)$$

où  $s^2 \equiv (n - k)^{-1} SSR(\hat{\beta})$ .

Le premier de ces estimateurs, l'expression (3.05), fait usage de l'estimateur  $\sigma^2$  du maximum de vraisemblance (consulter le Chapitre 8), qui est biaisé vers le bas. Pour se rendre compte de cela, remarquons que

$$SSR(\hat{\beta}) \leq SSR(\beta_0).$$

car  $SSR(\hat{\beta})$  minimise  $SSR(\beta)$ . De plus,

$$E(SSR(\beta_0)) = n\sigma_0^2,$$

sous le DGP présumé, car  $SSR(\beta_0)$  est alors tout simplement  $\sum_{t=1}^n u_t^2$ , et  $u_t^2$  a une espérance égale à  $\sigma_0^2$ . Ainsi, en faisant usage de  $\hat{\sigma}^2$  dans (3.05), nous tendrons à sous-estimer  $\sigma^2$  et donc, à sous-estimer la variabilité de nos estimations des paramètres.

Cela suggère que nous devrions prendre  $s^2$  au lieu de  $\hat{\sigma}^2$  lorsque nous estimons la matrice de covariance pour les estimations des paramètres dans les modèles de régression non linéaire, et ce en dépit du fait qu'il n'y ait aucune justification exacte dans le cas d'un échantillon fini pour pratiquer de la sorte. Il semble qu'un consensus autour de cette approche se manifeste, bien que quelques progiciels de régression non linéaire fassent toujours usage de  $\hat{\sigma}^2$ . La raison pour laquelle on utilise  $s^2$  est que dans le cas de la régression linéaire, il entraîne une estimation non biaisée de  $\sigma^2$ , et le bon sens (combiné à l'évidence apportée par les expériences Monte Carlo) suggère qu'il entraînera systématiquement une estimation moins biaisée dans le cas non linéaire également.

Le fait que  $s^2$  offre une estimation non biaisée de  $\sigma^2$  dans le cas de la régression non linéaire est sans aucun doute bien connu des lecteurs. Malgré tout, il reste suffisamment important pour que nous esquissons la démonstration. Dans le cas linéaire pour lequel  $\mathbf{x}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ ,

$$\begin{aligned} SSR(\hat{\boldsymbol{\beta}}) &\equiv (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y})^\top (\mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}) \\ &= \mathbf{y}^\top (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{y} \\ &= \mathbf{y}^\top \mathbf{M}_X \mathbf{y}. \end{aligned} \tag{3.07}$$

Sous le DGP que nous avons utilisé,  $\mathbf{y}^\top \mathbf{M}_X \mathbf{y}$  devient

$$(\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u})^\top \mathbf{M}_X (\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}).$$

Mais souvenons-nous que  $\mathbf{M}_X$  annule tout ce qui se situe dans  $\mathcal{S}(\mathbf{X})$ . L'espérance conditionnelle  $\mathbf{X}\boldsymbol{\beta}_0$  se situe assurément dans  $\mathcal{S}(\mathbf{X})$ , de sorte que  $SSR(\hat{\boldsymbol{\beta}})$  se résume à

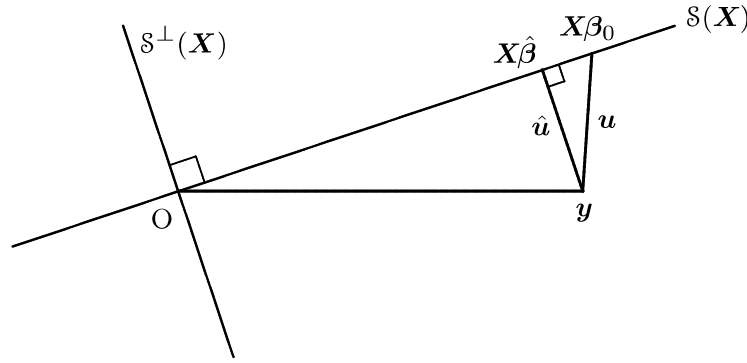
$$\mathbf{u}^\top \mathbf{M}_X \mathbf{u}.$$

L'espérance de cette expression est

$$\begin{aligned} E(\mathbf{u}^\top \mathbf{M}_X \mathbf{u}) &= E(\text{Tr}(\mathbf{u}^\top \mathbf{M}_X \mathbf{u})) \\ &= E(\text{Tr}(\mathbf{M}_X \mathbf{u} \mathbf{u}^\top)) \\ &= \text{Tr}(\mathbf{M}_X \sigma_0^2 \mathbf{I}) \\ &= \sigma_0^2 \text{Tr}(\mathbf{M}_X) \\ &= \sigma_0^2 (n - k), \end{aligned} \tag{3.08}$$

où la deuxième et la dernière ligne de (3.08) font usage de la propriété pratique de la trace que nous avons déjà rencontrée à la Section 1.6. Les lecteurs qui ne sont pas familiers avec ce résultat en (3.08) souhaiteront probablement consulter l'Annexe A.





**Figure 3.1** Les résidus sont inférieurs aux aléas

L'intuition de l'expression (3.08) est claire. Elle nous indique que, en moyenne, les résidus au carré sont  $(n - k)/n$  fois plus grands que les aléas originels au carré. De fait, alors, chacune des  $k$  dimensions de l'espace engendré par  $\mathbf{X}$  “absorbe” un des aléas d'origine. Cette “absorption” a été représentée sur la Figure 3.1 dans le cas d'un modèle de régression linéaire pour lequel  $k = 1$ . Ici  $\mathbf{y}$  est en fait égal à  $\mathbf{X}\beta_0 + \mathbf{u}$ . La longueur du vecteur  $\hat{\mathbf{u}} = \mathbf{M}_X \mathbf{y}$  est inférieure à la longueur du véritable vecteur d'aléas  $\mathbf{u}$ , car le premier est orthogonal à  $\mathcal{S}(\mathbf{X})$  alors que le second ne l'est pas.

Cette “absorption” des aléas originels engendre les erreurs que les moindres carrés entraînent en estimant le vecteur des coefficients. Les composantes de  $\mathbf{u}$  qui ne sont pas orthogonales à  $\mathbf{X}$  sont projetées sur  $\mathcal{S}(\mathbf{X})$  et donc finissent dans  $\hat{\beta}$ . Cela arrive à tous les éléments de  $\mathbf{u}$  à des degrés différents. Comme la discussion à propos de la pondération dans la Section 1.6 a dû le montrer, quelques résidus au carré seront moins de  $(n - k)/n$  fois plus grands que les aléas au carré correspondantes, alors que d'autres seront plus que  $(n - k)/n$  fois plus grandes. L'analyse algébrique nous montre que

$$\begin{aligned}
 E(\hat{u}_t^2) &= E(y_t - \mathbf{X}_t \hat{\beta})^2 \\
 &= E(y_t - \mathbf{X}_t \beta_0 + \mathbf{X}_t \beta_0 - \mathbf{X}_t \hat{\beta})^2 \\
 &= E(u_t - \mathbf{X}_t (\hat{\beta} - \beta_0))^2 \\
 &= E(u_t^2) - 2E(u_t \mathbf{X}_t (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}) \\
 &\quad + E(\mathbf{X}_t (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top) \\
 &= \sigma_0^2 - 2\sigma_0^2 \mathbf{X}_t (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top + \sigma_0^2 \mathbf{X}_t (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top \\
 &= \sigma_0^2 (1 - \mathbf{X}_t (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_t^\top) \\
 &= \sigma_0^2 M_{tt},
 \end{aligned}$$

où  $\mathbf{X}_t$  est la ligne  $t$  de  $\mathbf{X}$ . Ici,  $M_{tt}$ , l'élément diagonal  $t$  de  $\mathbf{M}_X$ , est égal à  $1 - h_t$ ,  $h_t$  étant l'élément diagonal  $t$  de  $\mathbf{P}_X$ . Ce résultat utilise le fait que  $E(u_t u_s) = 0$

pour tout  $t \neq s$ . Pour les observations fortement pondérées (c'est-à-dire avec un  $h_t$  fort), l'espérance du résidu au carré  $\hat{u}_t^2$  sera substantiellement inférieure à  $((n - k)/n)\sigma_0^2$ .

A partir de (3.08), il est clair que l'estimateur

$$s^2 \equiv \frac{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}}{n - k} \quad (3.09)$$

aura une espérance égale à  $\sigma_0^2$  si les données avaient été générées par un cas particulier de (3.02). Cet estimateur paraît donc être un estimateur raisonnable dans un contexte de moindres carrés ordinaires, et c'est ce que presque tous les programmes de régression par moindres carrés ordinaires utilisent lorsqu'ils calculent la matrice de covariance OLS. Mais notons que bien que  $s^2$  soit non biaisé pour  $\sigma^2$ ,  $s$  n'est pas non biaisé pour  $\sigma$ , car prendre la racine carrée de  $s^2$  n'est pas une opération linéaire.

Ces résultats doivent rendre évident le fait que dans le cas des modèles de régression linéaire, l'estimateur standard de la covariance OLS

$$\mathbf{V}_s(\hat{\boldsymbol{\beta}}) \equiv s^2 (\mathbf{X}^\top \mathbf{X})^{-1}, \quad (3.10)$$

offre une estimation non biaisée de la véritable matrice de covariance, qui est  $\sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ . Toutefois, et par contraste avec le cas linéaire, toute tentative de pratiquer des inférences à partir des estimations des modèles de régression non linéaire sera entravée par le fait que (3.04) n'est elle-même qu'une approximation, et que (3.06) est au plus une évaluation de (3.04). Malgré cela, la plupart des utilisateurs traitent (3.06), et même (3.05), exactement de la même manière que la matrice habituelle de covariance des OLS (3.10) pour construire des intervalles de confiance et des tests d'hypothèses pour de nombreux éléments de  $\boldsymbol{\beta}$ . Parce que (3.06) n'est qu'une évaluation, c'est pourtant une opération risquée. Malgré tout, dans les deux sections qui suivent, nous discuterons de ce qu'implique cette opération.

### 3.3 INTERVALLES DE CONFIANCE ET RÉGIONS DE CONFIANCE

Un **intervalle de confiance** pour un paramètre unique à un niveau quelconque  $\alpha$  (compris entre 0 et 1) est un intervalle sur la droite réelle construit de telle manière que l'on peut être sûr que la véritable valeur du paramètre se situera dans cet intervalle  $(1 - \alpha)\%$  du temps. Une **région de confiance** est conceptuellement la même chose, sauf que c'est une région dans un espace à  $l$  dimensions (habituellement l'analogie d'une ellipse à  $l$  dimensions) qui est construite de manière à ce que la vraie valeur d'un vecteur de paramètres à  $l$  composantes se trouve dans cette région  $(1 - \alpha)\%$  du temps. Remarquons que lorsque nous trouvons un intervalle de confiance, nous ne faisons aucune affirmation concernant la distribution du paramètre lui-même, mais plutôt

sur la probabilité que notre intervalle aléatoire, à cause de sa construction en termes des estimations des paramètres et de leur matrice de covariance, comprenne la véritable valeur.

Dans le contexte des modèles de régression, la construction d'un intervalle de confiance se fait généralement à l'aide d'une estimation du seul paramètre en cause, d'une estimation de son écart type, et d'une **valeur critique** que l'on prend soit dans la distribution normale, soit dans la distribution de Student. L'écart type estimé est bien évidemment la racine carrée de l'élément approprié de la diagonale de la matrice de covariance estimée. La valeur critique dépend de  $1 - \alpha$ , la probabilité que l'intervalle de confiance prenne la vraie valeur; si nous désirons que cette probabilité soit proche de un, la valeur critique devra être relativement importante, et par là l'intervalle de confiance également.

Supposons que le paramètre qui nous intéresse soit  $\beta_1$ , que son estimation NLS soit  $\hat{\beta}_1$ , et que l'écart type estimé de l'estimateur soit

$$\hat{S}(\hat{\beta}_1) \equiv s((\hat{\mathbf{X}}^\top \hat{\mathbf{X}})_{11})^{-1/2}.$$

Il nous faut tout d'abord connaître la longueur de notre intervalle de confiance en termes de l'écart type estimé  $\hat{S}(\hat{\beta}_1)$ . Nous recherchons donc la valeur de  $\alpha$  dans une table bilatérale de la distribution normale ou de la distribution de Student, ou la valeur  $\alpha/2$  dans une table unilatérale.<sup>1</sup> Cela nous donne la valeur critique  $c_\alpha$ . Nous trouvons donc un intervalle de confiance approximé allant de

$$\hat{\beta}_1 - c_\alpha \hat{S}(\hat{\beta}_1) \quad \text{à} \quad \hat{\beta}_1 + c_\alpha \hat{S}(\hat{\beta}_1), \quad (3.11)$$

qui comprendra la vraie valeur de  $\beta_1$ , grossièrement, dans  $(1-\alpha)\%$  des cas. Par exemple, si  $\alpha$  était .05 et si l'on utilisait les tables pour la distribution normale, nous trouverions une valeur critique bilatérale égale à 1.96. Cela implique que, pour la distribution normale avec une espérance  $\mu$  et une variance  $\omega^2$ , 95% de la masse de probabilité de cette distribution se situe entre  $\mu - 1.96\omega$  et  $\mu + 1.96\omega$ . Donc dans ce cas, notre intervalle de confiance approximé irait de

$$\hat{\beta}_1 - 1.96\hat{S}(\hat{\beta}_1) \quad \text{à} \quad \hat{\beta}_1 + 1.96\hat{S}(\hat{\beta}_1).$$

A l'évidence, nous faisons de fortes hypothèses lorsque nous construisons un intervalle de confiance de cette façon. Premièrement, il nous faut supposer que  $\hat{\beta}_1$  est normalement distribué, une hypothèse qui n'est strictement justifiée que si nous traitons un modèle de régression avec des régresseurs fixes et

<sup>1</sup> En réalité, nous laisserions sans doute l'ordinateur moderne remplir cette tâche. Un programme de statistiques convenable devrait être capable de donner la valeur critique associée à tout niveau de signification  $\alpha$ , ainsi que le niveau de signification correspondant à une valeur critique, pour les distributions normales, de Student, de Fisher et du  $\chi^2$ .

des aléas normaux.<sup>2</sup> Deuxièmement, nous supposons que  $\hat{S}(\hat{\beta}_1)$  est la véritable déviation moyenne de  $\hat{\beta}_1$ , ce qui ne sera en fait jamais le cas. A moins que le DGP ne soit un cas particulier du modèle que nous avons estimé, notre estimation de la matrice de covariance de  $\hat{\beta}$ , (3.06), ne sera généralement pas valide, pas même en tant qu'approximation. Dans le cas où elle serait correcte,  $s^2$  n'est qu'une estimation de  $\sigma^2$  et  $\hat{\mathbf{X}}^\top \hat{\mathbf{X}}$  n'est qu'une estimation de  $\mathbf{X}^\top(\beta_0)\mathbf{X}(\beta_0)$ , de sorte que  $\hat{S}(\hat{\beta}_1)$  sera probablement une estimation peu satisfaisante de  $S(\hat{\beta}_1)$ .

Dans le cas linéaire, il est traditionnel de ne traiter qu'un (et un seul!) de ces problèmes, c'est à dire le problème de la nature de  $s^2$  (qui est une estimation de  $\sigma^2$ ). Comme nous le montrerons plus tard dans la Section 3.5, pour les modèles de régression linéaire avec des régresseurs fixes et des aléas normaux, la quantité

$$\frac{\hat{\beta}_i - \beta_{0i}}{\hat{S}(\hat{\beta}_i)} \quad (3.12)$$

est distribuée selon la loi de Student à  $n - k$  degrés de liberté lorsque le DGP est un cas particulier du modèle que l'on estime. Ainsi, en prenant les valeurs critiques utilisées dans (3.11) à partir de la loi de Student à  $n - k$  degrés de liberté au lieu de la normale centrée réduite, nous trouvons l'intervalle de confiance exact dans ce cas particulier. En utilisant la distribution de Student, on intègre systématiquement le fait que  $s$  est un estimateur biaisé.

Lorsque l'on traite des modèles non linéaires, il est généralement plus fiable d'utiliser la distribution de Student à  $n - k$  degrés de liberté plutôt que la distribution normale standard. L'intervalle de confiance qui en résulte sera légèrement plus large, mais dans de nombreux cas, il ne sera toujours pas assez large. La plupart du temps (mais pas toujours), les problèmes mentionnés plus haut, résultent de l'étroitesse des intervalles de confiance estimés, de sorte qu'il est bon d'opérer mentalement une réduction du niveau de confiance associé à tout intervalle de ce genre. Ceci est particulièrement important lorsque le modèle a un fort degré de non linéarité, lorsque les aléas peuvent être très sensiblement non normaux, et lorsque la taille de l'échantillon est faible. Malheureusement, il n'existe pas de règle de sélection aisée qui nous dise de combien réduire le niveau de confiance dans la plupart des cas. Tout ce que nous pouvons faire généralement, c'est de se ramener à l'expérience, à l'évidence à partir des simulations Monte Carlo, et au bon sens.

Lorsque l'on s'intéresse à deux ou trois paramètres, il peut être très trompeur d'observer les intervalles de confiance pour les paramètres pris individuellement plutôt que la région de confiance de tous les paramètres pris conjointement. Pour s'en rendre compte, il est nécessaire de comprendre

<sup>2</sup> Cela se justifie parce que  $\hat{\beta} - \beta_0 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}$ , ce qui implique que  $\hat{\beta} - \beta_0$  est simplement une combinaison linéaire des variables aléatoires  $\mathbf{u}$  distribuées normalement, et doit donc être, lui-aussi, distribué normalement.

pourquoi une région de confiance pour  $l$  paramètres prend la forme d'une figure à  $l$  dimensions comparable à une ellipse. L'un des résultats présentés dans l'Annexe B est que, si  $\mathbf{x}$  est un vecteur à  $l$  composantes distribué normalement, à espérance nulle et à matrice de covariance  $\mathbf{V}$  non singulière et de dimension  $l \times l$ , alors la variable aléatoire scalaire donnée par la forme quadratique  $\mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x}$  suit une distribution du  $\chi^2$  à  $l$  degrés de liberté. Il nous est possible de construire une région de confiance pour tout sous-ensemble des composantes de  $\boldsymbol{\beta}$  en faisant usage de ce résultat.

Supposons que l'on désire construire une région de confiance pour les  $l$  premiers éléments du vecteur  $\boldsymbol{\beta}$  à  $k$  éléments où  $l > 1$ . Pour cette opération, nous aurons besoin d'une estimation de la matrice de covariance des  $l$  premiers éléments de  $\boldsymbol{\beta}$ . Si  $l = k$ , nous pouvons utiliser soit  $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$  soit  $\mathbf{V}_s(\hat{\boldsymbol{\beta}})$ , telles que nous les avons données dans (3.05) et (3.06). Si  $l < k$ , il nous faut utiliser une sous-matrice de dimension  $l \times l$  d'une de ces deux matrices de covariance estimées. Cette sous-matrice peut être obtenue par l'utilisation d'une formule sur l'inversion des matrices partitionnées (consulter l'Annexe A) ou, de façon plus simple, par l'utilisation du Théorème FWL.<sup>3</sup> Si l'on partitionne le vecteur complet de paramètres  $\boldsymbol{\beta}$  en  $[\boldsymbol{\beta}_1 : \boldsymbol{\beta}_2]$ , avec  $\boldsymbol{\beta}_1$  qui représente le sous-vecteur qui nous intéresse, et si l'on prend  $\hat{\sigma}^2$  comme estimateur de  $\sigma^2$ , nous obtenons

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_1) = \hat{\sigma}^2 (\hat{\mathbf{X}}_1^\top \hat{\mathbf{M}}_2 \hat{\mathbf{X}}_1)^{-1},$$

où  $\hat{\mathbf{M}}_2$  projette sur le complémentaire de  $\mathcal{S}(\hat{\mathbf{X}}_2)$ . Il est ainsi aussi facile de traiter le cas où  $l < k$  que le cas où  $l = k$ . Pour conserver la simplicité de la notation toutefois, nous supposerons pour la suite de notre discussion que nous construisons une région de confiance pour la totalité des paramètres du vecteur  $\boldsymbol{\beta}$ , de sorte que  $l = k$ . Par souci de réalisme, nous supposerons également que la matrice de covariance estimée de  $\hat{\boldsymbol{\beta}}$  est  $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$ , bien qu'elle puisse être aussi bien  $\mathbf{V}_s(\hat{\boldsymbol{\beta}})$ .

Notons  $\boldsymbol{\beta}_0$  la véritable (et inconnue) valeur de  $\boldsymbol{\beta}$ . Considérons la forme quadratique

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\beta}}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \quad (3.13)$$

qui est une variable aléatoire scalaire dépendant du vecteur aléatoire  $\hat{\boldsymbol{\beta}}$ . Elle n'aura en fait une distribution du  $\chi^2$  à  $l$  degrés de liberté pour des échantillons de taille finie ni pour une régression linéaire, ni pour une régression non linéaire. Mais il est raisonnable d'espérer qu'elle suivra approximativement une distribution du  $\chi^2(l)$ , et en réalité une telle approximation est correcte lorsque la taille de l'échantillon est suffisamment élevée; voir la Section 5.7. Par conséquent, avec autant de justifications (ou peut-être avec aussi peu!)

<sup>3</sup> Bien que l'usage du Théorème FWL soit incontestable ici, il n'est peut-être pas évident qu'il soit approprié lorsque le modèle est non linéaire. Dans le cas non linéaire, il devrait être appliqué à la régression de Gauss-Newton, dont nous discuterons au Chapitre 6.

que pour le cas d'un paramètre unique, la région de confiance pour  $\beta$  est élaborée comme si (3.13) suivait réellement une distribution du  $\chi^2(l)$ .<sup>4</sup>

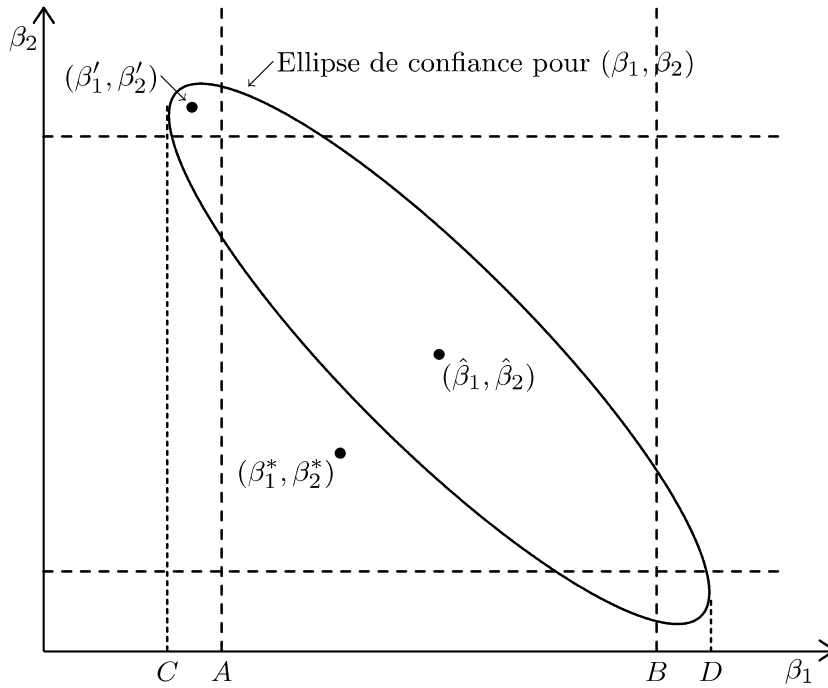
Pour un ensemble donné d'estimations  $\hat{\beta}$ , la région de confiance (approximée) à un niveau de confiance  $\alpha$  peut être définie comme l'ensemble des vecteurs  $\beta_0$  pour lesquels la valeur de (3.13) est inférieure à une valeur critique donnée, disons  $c_\alpha(l)$ . Cette valeur critique sera telle que, si  $z$  est une variable aléatoire suivant une distribution du  $\chi^2(l)$ ,

$$\Pr(z > c_\alpha(l)) = \alpha.$$

La région de confiance est donc l'ensemble de tous les  $\beta$  pour lesquels

$$(\hat{\beta} - \beta)^\top \hat{V}^{-1}(\hat{\beta})(\hat{\beta} - \beta) < c_\alpha(l). \quad (3.14)$$

Puisque le membre de gauche de cette équation est quadratique en  $\beta$ , la région est, pour  $l = 2$ , l'intérieur d'une ellipse, et pour  $l > 2$ , l'intérieur d'une ellipsoïde à  $l$  dimensions.



**Figure 3.2** Ellipse de confiance et intervalles de confiance

<sup>4</sup> Il est également envisageable de construire une région de confiance estimée en utilisant la loi du  $F(l, n - k)$ , ce qui pourrait éventuellement fournir une meilleure approximation dans les échantillons finis. On se sert ici du  $\chi^2$  pour simplifier l'exposé.

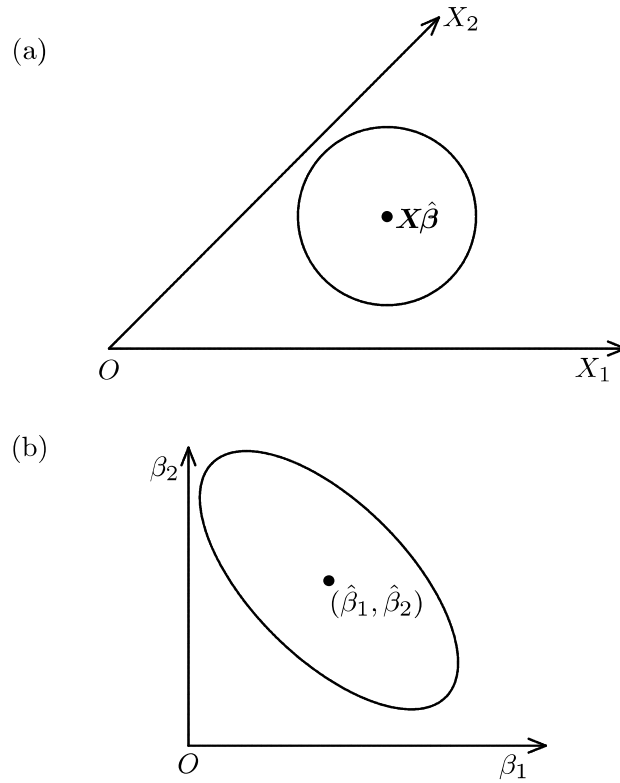
La Figure 3.2 illustre ce à quoi peut ressembler une ellipse de confiance dans un cas bi-dimensionnel. Dans ce cas, les deux estimations des paramètres sont corrélées négativement, et sont centrées autour des estimations  $(\hat{\beta}_1, \hat{\beta}_2)$  des paramètres. Les intervalles de confiance pour  $\beta_1$  et  $\beta_2$  sont aussi représentés; et il devrait être clair désormais qu'il peut être trompeur de ne considérer que ces derniers plutôt que l'ellipse de confiance. D'une part, il y a, à l'évidence, quelques points tels que  $(\beta_1^*, \beta_2^*)$ , qui se situent à l'extérieur de l'ellipse de confiance et à l'intérieur des intervalles de confiance, et d'autre part, il y a des points comme  $(\beta_1', \beta_2')$ , qui appartiennent à l'ellipse mais qui se situent à l'extérieur d'une ou de deux régions de confiance.

Il est très utile de passer davantage de temps sur la Figure 3.2, pour voir plus précisément la relation entre l'ellipse de confiance et les intervalles de confiance unidimensionnels. Il est tentant de penser que ces derniers devraient être obtenus par les points correspondant aux extrémités de l'ellipse de confiance, de sorte que, par exemple, l'intervalle de confiance pour  $\beta_1$  dans la figure serait donné par le segment de droite  $CD$ . Cela est cependant incorrect, ce qui peut être vu de deux façons différentes et révélatrices.

Le premier argument est le suivant. Le membre de droite de (3.14) est une valeur critique pour une distribution du  $\chi^2$  avec, dans le cas de notre figure, deux degrés de liberté. Si l'on porte son attention sur l'intervalle de confiance d'un paramètre unique, la distribution du  $\chi^2$  qui est pertinente n'aurait qu'un seul degré de liberté. Pour un niveau de confiance donné  $\alpha$ , la valeur critique est une fonction croissante du nombre de degrés de liberté. Dans le cas présent, la valeur critique à 5% pour une variable suivant une distribution du  $\chi^2$  à un degré de liberté est 3.84, et pour deux degrés de liberté, elle est égale à 5.99.

Le second argument est plus général. Souvenons-nous que la région de confiance est définie de telle sorte qu'elle contienne les véritables valeurs des paramètres avec une probabilité de  $1 - \alpha$ . Mais il est possible d'inverser les rôles de  $\beta$  et  $\hat{\beta}$ . Si les vraies valeurs des paramètres étaient données par  $\beta$ , la région définie par (3.14) serait une région dans laquelle la variable aléatoire  $\hat{\beta}$  serait réalisée avec la probabilité  $1 - \alpha$ . Ainsi l'ellipse de confiance contient une masse de probabilité de  $1 - \alpha$ . Il en va de même pour l'intervalle de confiance de  $\beta_1$ : il doit également contenir une masse de probabilité de  $1 - \alpha$ . Dans la construction bi-dimensionnelle de la Figure 3.2, le rectangle infini limité par les lignes verticales passant par les points  $A$  et  $B$  doit avoir cette masse de probabilité, puisque nous voulons permettre à  $\beta_2$  de prendre n'importe quelle valeur réelle. Parce que le rectangle infini et l'ellipse de confiance doivent avoir la même masse de probabilité, aucun ne peut contenir l'autre, et nous voyons pourquoi l'ellipse doit déborder de la région définie par l'intervalle de confiance unidimensionnel.

Il est clair d'après (3.13) que l'orientation de l'ellipse de confiance et les longueurs relatives de ses axes sont déterminées par la matrice de covariance estimée  $\hat{V}(\hat{\beta})$ . Si cette dernière était diagonale, les axes de l'ellipse seraient



**Figure 3.3** Ellipses de confiance de forme elliptique

parallèles aux axes des coordonnées. Et si tous les éléments diagonaux étaient égaux, la région de confiance serait une sphère. Il y a, toutefois, un autre moyen de représenter géométriquement une région de confiance, un moyen avec lequel elle prendra toujours une forme analogue à une sphère à  $l$  dimensions. Cette représentation est assez révélatrice. Par souci de simplicité, nous nous limiterons au cas des modèles de régression linéaire avec  $n$  observations et deux paramètres,  $\beta_1$  et  $\beta_2$ .

Considérons l'espace à  $n$  dimensions dans lequel les variables sont représentées par des vecteurs. Maintenant, portons notre attention sur le sous-ensemble bi-dimensionnel de l'espace à  $n$  dimensions engendré par les deux vecteurs  $\mathbf{X}_1$  et  $\mathbf{X}_2$ . Cet espace à deux dimensions est dessiné dans la partie (a) de la Figure 3.3, dans laquelle nous voyons également le vecteur de valeurs ajustées,  $\mathbf{X}\hat{\beta}$ . Nous avançons que le cercle tracé autour de  $\mathbf{X}\hat{\beta}$  et de rayon  $\hat{\sigma}\sqrt{c_\alpha(2)}$  correspond à l'ellipse de confiance de  $\beta_1$  et  $\beta_2$ . En réalité, il est aisé de saisir cela. L'équation du cercle est

$$\|\mathbf{y} - \mathbf{X}\hat{\beta}\| = \hat{\sigma}\sqrt{c_\alpha(2)}. \quad (3.15)$$

Tout vecteur  $\mathbf{y}$  appartenant à  $\mathcal{S}(\mathbf{X})$  peut s'exprimer par  $\mathbf{X}\beta$  pour un  $\beta$  donné. Ainsi (3.15) peut s'écrire

$$\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\| = \hat{\sigma}\sqrt{c_\alpha(2)}$$



ou, en prenant les carrés,

$$(\hat{\beta} - \beta)^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \beta) = \hat{\sigma}^2 c_\alpha(2). \quad (3.16)$$

Puisque l'estimation de la matrice de covariance de  $\hat{\beta}$  est  $\hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$ , (3.16) est simplement l'équation qui délimite l'ellipse de confiance pour  $\beta$ ; à comparer avec (3.14).

Cela signifie que nous disposons de deux manières de représenter géométriquement l'espace dans lequel se situe  $\beta$ . L'une est la représentation directe illustrée dans la Figure 3.2 et dans la partie (b) de la Figure 3.3, pour laquelle il y a deux axes mutuellement perpendiculaires pour les directions des deux paramètres  $\beta_1$  et  $\beta_2$ , et pour laquelle la région de confiance ressemble à une ellipse. L'autre est illustrée dans la partie (a) de la Figure 3.3, dans laquelle un point  $\beta$  est représenté par le vecteur  $\mathbf{X}\beta$ . Dans un certain sens, la seconde représentation est la plus naturelle, parce que la région de confiance y est symétrique dans toutes les directions. Bien entendu, cette symétrie dépend de l'hypothèse selon laquelle tous les aléas ont même variance,  $\sigma^2$ , et nous ne l'obtiendrions pas si  $E(u_t^2 | \mathbf{X}_t)$  était une fonction de  $\mathbf{X}_t$ . Le point crucial est que le cercle de la partie (a) et l'ellipse de la partie (b) contiennent tous deux la même région de confiance.

La brève discussion que nous avons tenue à propos des intervalles de confiance et des régions de confiance reposait sur des arguments géométriques. Pour une approche plus traditionnelle, les lecteurs intéressés pourront consulter une référence telle que l'ouvrage de Kendall et Stuart (1979, Chapitre 20). Elle offre une discussion plus détaillée sur la signification des intervalles de confiance et sur la manière de bâtir différents types d'intervalles et de régions de confiance lorsque la distribution d'échantillonnage des estimations des paramètres est connue. Le problème attaché aux modèles de régression non linéaire réside dans le fait que la distribution d'échantillonnage n'est jamais connue avec exactitude. De ce fait, à moins que l'échantillon ne soit très important, il n'y a souvent pas d'intérêt à essayer de construire des formes sophistiquées d'intervalles de confiance et de régions de confiance.

### 3.4 TESTS D'HYPOTHÈSES: INTRODUCTION

Les économistes désirent souvent tester les hypothèses des modèles de régression qu'ils estiment. De telles hypothèses prennent généralement la forme de contraintes d'égalité sur quelques uns des paramètres. Elles peuvent impliquer des tests pour savoir si un paramètre unique prend une certaine valeur (disons  $\beta_2 = 1$ ), si deux paramètres sont reliés de telle ou telle façon (disons  $\beta_3 = 2\beta_4$ ), si une restriction non linéaire telle que  $\beta_1/\beta_3 = \beta_2/\beta_4$  est valide, ou si un ensemble complet de contraintes linéaires et/ou non linéaires est acceptable. L'hypothèse selon laquelle la restriction ou l'ensemble des restrictions que l'on désire tester est valide, est appelée **hypothèse nulle**, et on la note souvent

$H_0$ . Le modèle pour lequel les contraintes sont inexactes est généralement appelé **hypothèse alternative**, ou quelquefois **hypothèse maintenue**, et est noté  $H_1$ . La terminologie “hypothèse maintenue” provient du fait que lors d'un test statistique seule l'hypothèse nulle  $H_0$  est soumise au test. Rejeter  $H_0$  n'oblige en rien à accepter  $H_1$ , puisque ce n'est pas  $H_1$  que l'on soumet au test. Considérons ce qu'il adviendrait si le DGP n'était pas un cas particulier de  $H_1$ . Clairement  $H_0$  et  $H_1$  seraient simultanément fausses, et il est fort possible qu'un test sur  $H_0$  conduise à son rejet. D'autres tests pourraient permettre de rejeter l'hypothèse fausse  $H_1$ , mais uniquement si elle prenait la place de l'hypothèse nulle et si l'on formulait de nouvelles hypothèses alternatives.

Tous les tests d'hypothèses discutés dans cet ouvrage impliquent l'élaboration de **statistiques de test**. Une statistique de test, disons  $T$ , est une variable aléatoire dont la distribution de probabilité est connue, soit avec exactitude, soit approximativement, sous l'hypothèse nulle. Nous examinons ensuite avec quelle probabilité la valeur observée de  $T$  peut survenir, en fonction de cette distribution de probabilité. Si  $T$  est un nombre qui a pu survenir avec une grande probabilité, alors nous n'avons aucune raison de rejeter l'hypothèse nulle  $H_0$ . Toutefois, si c'est un nombre qui ne surviendrait qu'avec une faible probabilité, il faut nous rendre à l'évidence, et la rejeter.

La meilleure façon d'exécuter un test est de diviser l'ensemble de toutes les valeurs possibles de  $T$  en deux régions, la **région d'acceptation** et la **région de rejet** (ou **région critique**). Si la valeur de  $T$  appartient à la région d'acceptation, on accepte l'hypothèse nulle (ou du moins on ne la rejette pas, au taux choisi) alors que si elle appartient à la région de rejet, on la rejette.<sup>5</sup> Par exemple, si l'on savait que  $T$  suit une distribution du  $\chi^2$ , la région d'acceptation serait composée de toutes les valeurs de  $T$  égales ou inférieures à une certaine **valeur critique**, disons  $C$ . Si au lieu de cela,  $T$  était telle qu'elle suive une distribution normale, alors pour un test bilatéral, la région d'acceptation serait formée par toutes les valeurs absolues de  $T$  inférieures ou égales à  $C$ , de sorte que la région de rejet serait composée de deux parties, l'une contenant les valeurs supérieures à  $C$  et l'autre contenant les valeurs inférieures à  $-C$ .

Le **niveau de signification** ou plus simplement le **niveau** d'un test est la probabilité que la statistique de test rejettera l'hypothèse nulle lorsque celle-ci est vraie. Notons  $\theta$  le vecteur de paramètres que l'on doit tester,  $\Theta_0$  l'ensemble des valeurs de  $\theta$  qui satisfont  $H_0$ , et  $R$  la région de rejet. Alors le niveau de

<sup>5</sup> Les termes “région d'acceptation” et “région de rejet” font également référence à des sous-échantillons de l'espace d'échantillonnage. Tout échantillon, disons  $y$ , génère une statistique de test  $T$ . Si la valeur de  $T$  appartient à la région de rejet, alors  $y$  doit aussi y appartenir, et inversement si la valeur de  $T$  appartient à la région d'acceptation. Dans ce sens, l'échantillon tout entier peut être partitionné en deux régions qui correspondent à la région d'acceptation et à la région de rejet de l'hypothèse nulle.

la statistique de test  $T$  est

$$\alpha \equiv \Pr(T \in R \mid \theta \in \Theta_0).$$

Conventionnellement, le niveau est choisi à une valeur très faible, généralement de l'ordre de .001 à .10. Il est choisi de façon plus ou moins arbitraire, et comme nous le verrons au cours d'une discussion ultérieure, cette caractéristique des tests d'hypothèses est quelquefois assez peu satisfaisante.

Nous exécutons des tests dans l'espoir de les voir rejeter l'hypothèse nulle lorsqu'elle est inexacte. En conséquence, la **puissance** du test est d'un grand intérêt. La puissance d'une statistique de test  $T$  est la probabilité que  $T$  rejette l'hypothèse nulle lorsque celle-ci est fausse. De façon formelle, elle peut se définir comme

$$\Pr(T \in R \mid \theta \notin \Theta_0).$$

La puissance dépendra évidemment de la façon dont les données ont été générées. Si l'hypothèse nulle est seulement légèrement fausse, nous espérons que la puissance sera inférieure à ce qu'elle serait si l'hypothèse nulle était manifestement fausse. Nous nous attendons également à ce que la puissance augmente avec la taille de l'échantillon,  $n$ . Si pour tout  $\theta$  dans une région donnée de l'espace des paramètres, disons  $\Theta_1$ , la puissance du test tend vers un lorsque  $n \rightarrow \infty$ , le test est dit **convergent** contre les hypothèses alternatives dans  $\Theta_1$ . Bien sûr, un test peut être convergent contre certaines hypothèses alternatives et pas contre d'autres. Nous discuterons au cours des Chapitres 12 et 13 de ce qui détermine exactement la puissance des statistiques de test, lorsque le DGP est, et n'est pas, un cas particulier de l'hypothèse alternative.

La façon traditionnelle d'exécuter un test est tout d'abord de choisir son niveau, et ensuite d'utiliser ce niveau pour déterminer une valeur critique que l'on trouve dans les tables de distribution appropriée. Par exemple, si une statistique de test est distribuée suivant la  $\chi^2(1)$  sous l'hypothèse nulle, la valeur critique à .05 (ou 5%) est 3.84, car la probabilité d'obtenir un tirage aléatoire à partir de la  $\chi^2(1)$  qui soit supérieur à 3.84 est de .05. Puis, si la statistique de test s'avérait être, disons 3.51, nous ne rejeterions pas l'hypothèse nulle au niveau .05, alors que si elle se révélait être supérieure, disons par exemple 5.43, nous devrions rejeter l'hypothèse nulle à ce niveau de .05.

Il y a deux problèmes relatifs à cette procédure. Premièrement, le choix du niveau du test est plus ou moins arbitraire. Cela reflète l'intensité de notre désir de faire une erreur en rejetant l'hypothèse nulle lorsqu'elle est exacte (ou de commettre l'**erreur de première espèce**), plutôt que d'accepter l'hypothèse nulle lorsqu'elle est fausse (ou de commettre l'**erreur de deuxième espèce**). Si nous désirons fortement éviter une erreur de première espèce, nous utiliserons un très faible niveau de signification, certainement inférieur à .01 et probablement .001, ou même moins encore. Si nous sommes plus sensibles à l'erreur de deuxième espèce, nous prendrons un niveau de signification plus fort, comme

.05 ou même .10. Si un utilisateur exécute un test au niveau .05 et qu'un autre décide de l'exécuter au niveau .01, ils aboutiront sans aucun doute à des résultats divergents. Cela pose de sérieux problèmes pour les lecteurs qui tentent d'interpréter ces résultats, particulièrement si seules les conclusions des tests (plutôt que les vraies valeurs des statistiques de test) ont été mentionnées. Ne rapporter que les conclusions des tests est une pratique que nous abhorrons, et que nous déconseillons plus que vivement.

Cela nous amène à considérer le second problème. Si un auteur rapporte en fait les valeurs des statistiques de test, les lecteurs peuvent potentiellement établir quelques conclusions, mais cela leur demandera des efforts pour convertir les statistiques de test en nombres signifiant quelque chose d'intéressant. Cela devient plus facile pour le lecteur lorsque l'auteur adopte un moyen alternatif de présenter les résultats des tests, un moyen qui se popularise désormais. Cela consiste à mentionner le ***P* marginal** (ou **niveau marginal de signification**) associé à toute statistique de test, soit avec, soit à la place des statistiques elles-mêmes, puisque ces dernières ne comprennent pas plus d'information que les *P* marginaux. Le *P* marginal est la probabilité d'observer, si la statistique de test était réellement distribuée comme elle le serait sous l'hypothèse nulle, une valeur de la statistique de test au moins aussi extrême que celle observée. Si elle est inférieure à  $\alpha$ , alors nous rejeterions  $H_0$  au niveau  $\alpha$ .

Il n'est généralement pas possible de calculer les *P* marginaux à partir des tables de distribution, mais cela devient envisageable si l'on dispose d'un logiciel informatique adéquat (qui devrait être inclus dans tout progiciel de régression moderne); l'usage croissant d'ordinateurs toujours plus performants joue certainement dans la diffusion de l'approche de l'utilisation du *P* marginal. Dans l'exemple précédent, au lieu de rapporter les valeurs des statistiques de test 3.51 et 5.43, le chercheur devrait plutôt se servir du fait qu'elles sont supposées suivre une distribution du  $\chi^2(1)$  sous  $H_0$ , et rapporter les valeurs *P* marginaux respectifs de .0610 et .0198. Ces valeurs de *P* nous indiquent que pour la distribution du  $\chi^2(1)$ , les valeurs au moins aussi fortes que 3.51 apparaîtraient pour environ 6.1% dans les tirages, alors que les nombres supérieurs ou égaux à 5.43 n'apparaîtraient que pour moins de 2% dans les tirages.

Le *P* marginal ne nous oblige pas à établir une décision concernant l'hypothèse nulle. Si nous obtenons un *P* marginal, disons d'environ .000001, nous voudrions presque sûrement rejeter l'hypothèse nulle. Mais si nous obtenons un *P* marginal de .04 ou même de .004, rien ne nous oblige à la rejeter. Nous considérerions ce résultat comme une information qui nous permet d'émettre quelques doutes à propos de l'hypothèse nulle, mais non comme un résultat probant par lui-même. Nous voyons que cette approche quelque peu agnostique envers les statistiques de test, dans le sens où elles sont considérées au plus comme des indices à partir desquels on pourrait ou non prendre des décisions, est généralement la plus raisonnable à adopter. Elle est peut-être particulièrement appropriée dans le cas des modèles de régression non linéaire,

ou les  $P$  marginaux ne sont généralement que des approximations (et quelquefois assez imprécises!).

Il est crucial de comprendre qu'une valeur de  $P$  n'est pas la probabilité que l'hypothèse nulle soit correcte, et par elle-même ne nous permet pas de déduire cette probabilité. Dans l'élaboration traditionnelle des tests d'hypothèses, une hypothèse est soit vraie, soit fausse, et nous ne pouvons pas parler d'une probabilité quelconque qu'elle soit vraie. Dans la méthodologie Bayésienne, que nous n'utiliserons pas dans cet ouvrage, il est possible de dire qu'il y a une probabilité qu'une hypothèse soit vraie, mais il faut alors faire référence à une **information a priori** sur cette probabilité et il faut également préciser quelles sont les autres hypothèses qui pourraient être vraies. Puis étant donnée la probabilité a priori qu'une hypothèse soit vraie et les résultats de la statistique de test, il est possible de calculer une **probabilité a posteriori** que l'hypothèse soit correcte. Cette dernière probabilité sera sûrement plus importante que le  $P$  marginal.<sup>6</sup>

Comme nous l'avons précédemment indiqué, la distribution qu'une statistique de test est censée suivre sous l'hypothèse nulle peut être ou ne pas être connue avec précision. Si elle est connue, alors le test est communément appelé un **test exact**; dans le cas contraire, le test sera généralement basé sur une distribution dont on sait qu'elle s'applique uniquement asymptotiquement et s'appelle alors un **test asymptotique**. Les tests exacts sont rarement disponibles pour les modèles de régression non linéaire. Toutefois, il y a un cas bien particulier pour lequel les tests exacts de contraintes linéaires sont utilisables. La fonction de régression doit alors être linéaire, il doit être possible de traiter les régresseurs comme s'ils étaient **fixés dans les échantillons répétés**, et les aléas doivent être normalement et identiquement distribués. Dans la section qui suit, nous définirons le terme "fixés dans les échantillons répétés" et discuterons de ce cas particulier. Puis à la Section 3.6, nous aborderons les trois principes de base, soulignerons l'élaboration de la plupart des statistiques de test, et nous verrons comment elles pourraient être appliquées aux modèles de régression non linéaire.

### 3.5 TESTS D'HYPOTHÈSES DANS LES RÉGRESSIONS LINÉAIRES

Tous les étudiants en économétrie sont familiers avec les **Students** pour les tests d'hypothèses relatives à un paramètre unique et les **Fisher** pour les tests d'hypothèses concernant plusieurs paramètres simultanément. Si  $\hat{\beta}_i$  désigne l'estimation par moindres carrés du paramètre  $\beta_i$ , le Student pour tester l'hypothèse d'égalité de  $\beta_i$  à une quelconque valeur donnée  $\beta_{0i}$  est tout simplement égal à  $\hat{\beta}_i - \beta_{0i}$  divisée par une estimation de l'écart type de  $\hat{\beta}_i$  (expression (3.12)). Si  $\hat{\beta}$  désigne un ensemble d'estimations non contraintes

<sup>6</sup> Pour une discussion des relations entre le  $P$  marginal et l'inférence Bayésienne, consulter entre autre Lindley (1957), Shafer (1982), et Berger et Sellke (1987).

par moindres carrés et  $\tilde{\beta}$  désigne un ensemble d'estimations soumises à  $r$  contraintes distinctes, le Fisher pour le test de ces contraintes peut être calculée grâce à

$$\frac{(SSR(\tilde{\beta}) - SSR(\hat{\beta}))/r}{SSR(\hat{\beta})/(n - k)} = \frac{1}{rs^2} (SSR(\tilde{\beta}) - SSR(\hat{\beta})). \quad (3.17)$$

Les tests fondés sur les Students ou Fisher peuvent être aussi bien exacts qu'approximatifs. Dans le cas particulier dont nous avons discuté à la fin de la section précédente, dans lequel le modèle de régression et les contraintes sont linéaires en leurs paramètres, les régresseurs sont (ou peuvent être considérés comme étant) fixés dans les échantillons répétés, et les aléas sont normalement et indépendamment distribués, les Students et Fisher ordinaires suivent en échantillons finis, sous l'hypothèse nulle, la distribution dont ils portent le nom. Bien que ce cas ne soit pas aussi courant qu'on le voudrait, ces résultats sont suffisamment importants pour être traités dans une section séparée. De plus, il est utile de conserver à l'esprit le cas linéaire lorsque l'on considère le cas des modèles de régression non linéaire.

Considérons le modèle contraint

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u} \quad (3.18)$$

et le modèle non contraint

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \quad (3.19)$$

où les aléas  $\mathbf{u}$  sont supposés suivre une distribution normale avec un vecteur d'espérance nulle et une matrice de covariance  $\sigma^2\mathbf{I}$ . Le vecteur de paramètres  $\boldsymbol{\beta}$  a été scindé en deux sous-vecteurs,  $\boldsymbol{\beta}_1$  et  $\boldsymbol{\beta}_2$ , respectivement composés de  $k - r$  et  $r$  éléments. Nous voulons tester l'hypothèse de nullité de  $\boldsymbol{\beta}_2$ . Les estimations contraintes sont  $\tilde{\beta} = [\tilde{\beta}_1 : \mathbf{0}]$ , et les estimations non contraintes sont  $\hat{\beta} = [\hat{\beta}_1 : \hat{\beta}_2]$ . En limitant notre attention à des contraintes de nullité, nous ne limitons en rien la généralité des résultats puisque, ainsi que nous l'avons montré à la Section 1.3, tout ensemble de contraintes linéaires appliqué à un modèle linéaire peut toujours, par une reparamétrisation adéquate, être écrit comme un ensemble de contraintes de nullité.

Il suit de l'expression (3.07) que  $SSR(\tilde{\beta}) = \mathbf{y}^\top \mathbf{M}_1 \mathbf{y}$ , où  $\mathbf{M}_1$  désigne la projection sur  $\mathcal{S}^\perp(\mathbf{X}_1)$ , le complément orthogonal de l'espace engendré par  $\mathbf{X}_1$ . De façon similaire,  $SSR(\hat{\beta}) = \mathbf{y}^\top \mathbf{M}_X \mathbf{y}$ , où  $\mathbf{M}_X$  désigne la projection sur  $\mathcal{S}^\perp(\mathbf{X})$ , le complément orthogonal de l'espace engendré par  $\mathbf{X} \equiv [\mathbf{X}_1 \ \mathbf{X}_2]$ . Grâce au Théorème FWL (voir Section 1.4),  $\mathbf{y}^\top \mathbf{M}_X \mathbf{y}$ , qui est la somme des résidus au carré de la régression (3.19), est identique à la SSR de la régression

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \boldsymbol{\beta}_2 + \text{résidus}.$$

Ainsi nous constatons que

$$SSR(\hat{\beta}) \equiv \mathbf{y}^\top \mathbf{M}_X \mathbf{y} = \mathbf{y}^\top \mathbf{M}_1 \mathbf{y} - \mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}. \quad (3.20)$$

De (3.07) et (3.20), il vient que  $r$  fois le numérateur du Fisher de (3.17) est

$$\mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} = \|\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y}\|^2, \quad (3.21)$$

où  $\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \equiv \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1$  est la matrice projetant sur l'espace  $\mathcal{S}(\mathbf{M}_1 \mathbf{X}_2)$ . Sous le DGP

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_{10} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}), \quad (3.22)$$

les deux membres de l'expression (3.21) deviennent

$$\mathbf{u}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{u} = \|\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{u}\|^2, \quad (3.23)$$

car  $\mathbf{M}_1$  annule  $\mathbf{X}_1 \boldsymbol{\beta}_{10}$ .

L'expression (3.23), qui est égale à  $r$  fois le numérateur du Fisher sous le DGP (3.22), peut être considérée comme une forme quadratique du vecteur aléatoire à  $r$  composantes

$$\mathbf{v} \equiv \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{u}, \quad (3.24)$$

qui est distribué normalement (puisque ce n'est qu'une combinaison linéaire des variables aléatoires normalement distribuées que sont les éléments de  $\mathbf{u}$ ) avec une espérance nulle et une matrice de covariance

$$\boldsymbol{\Omega} \equiv E(\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{u} \mathbf{u}^\top \mathbf{M}_1 \mathbf{X}_2) = \sigma_0^2 \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2. \quad (3.25)$$

Ainsi, en combinant (3.24) et (3.25), le membre de droite de l'expression (3.23) devient

$$\sigma_0^2 \mathbf{v}^\top \boldsymbol{\Omega}^{-1} \mathbf{v}. \quad (3.26)$$

Cette expression est  $\sigma_0^2$  fois une forme quadratique en  $\mathbf{v}$  (vecteur à  $r$  composantes), qui est un vecteur normal multivarié d'espérance nulle et de matrice de covariance égale à l'inverse de la matrice de covariance de  $\mathbf{v}$ ,  $\boldsymbol{\Omega}$ . Ainsi, en faisant usage du résultat familier sur les formes quadratiques des vecteurs aléatoires normaux dont nous avons déjà parlé à la Section 3.3, nous en arrivons à la conclusion que  $1/\sigma_0^2$  fois la forme quadratique (3.26) est distribué selon le  $\chi^2(r)$ .

Si la taille de l'échantillon était infiniment grande, nous pourrions nous arrêter à cette conclusion. Le  $s^2$  estimé correspondrait sans distinction possible à la véritable valeur de  $\sigma_0^2$  de sorte que, sous l'hypothèse nulle,  $r$  fois le Fisher (3.17) serait égal à

$$\frac{1}{\sigma_0^2} \mathbf{u}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{u},$$

et nous avons précisément remarqué que cette quantité suit une distribution du  $\chi^2(r)$  sous  $H_0$ . Mais dans le cas d'un échantillon fini,  $s^2$  sera une variable

aléatoire qui évalue  $\sigma_0^2$  avec une précision qui croît avec  $n - k$ . Il nous faut, par conséquent, considérer la distribution du dénominateur du Fisher (3.17). Sous le DGP (3.22),  $(n - k)s^2$  est grâce à l'équation (3.09), égale à

$$\mathbf{u}^\top \mathbf{M}_X \mathbf{u}. \quad (3.27)$$

Dans l'Annexe B, nous démontrons que toute forme quadratique idempotente de variables aléatoires normales centrées réduites et indépendantes suit une distribution du  $\chi^2$  dont le nombre de degrés de liberté correspond au rang de la matrice idempotente. De façon évidente,  $\mathbf{u}^\top \mathbf{M}_X \mathbf{u} / \sigma_0^2$  est une telle forme quadratique idempotente, et puisque le rang de  $\mathbf{M}_X$  est  $n - k$ , nous en concluons que  $1/\sigma_0^2$  fois le dénominateur du Fisher suit une distribution du  $\chi^2(n - k)$ .

Ainsi le Fisher (3.17) est  $(n - k)/r$  fois le ratio de deux variables aléatoires, le numérateur étant distribué selon le  $\chi^2(r)$  et le dénominateur selon le  $\chi^2(n - k)$ . À condition que ces deux variables aléatoires soient mutuellement indépendantes, leur rapport sera distribué selon la  $F(r, n - k)$  (voir Annexe B). Une condition suffisante pour qu'elles soient indépendantes est

$$\mathbf{M}_X \mathbf{P}_{M_1 X_2} = \mathbf{0}.$$

Ce qui est en effet le cas, car  $\mathcal{S}(\mathbf{M}_1 \mathbf{X}_2)$ , l'espace engendré par  $\mathbf{M}_1 \mathbf{X}_2$ , est un sous-espace de  $\mathcal{S}(\mathbf{X})$ , l'espace engendré par  $\mathbf{X}_1$  et  $\mathbf{X}_2$ ; pour l'apercevoir, observons que  $\mathbf{M}_1 \mathbf{X}_2 = \mathbf{X}_2 - \mathbf{P}_1 \mathbf{X}_2$ . Ainsi  $\mathbf{M}_X$  annule  $\mathbf{P}_{M_1 X_2}$ .

Voici une explication plus intuitive de l'indépendance des formes quadratiques (3.23) et (3.27). La forme quadratique qui apparaît dans le dénominateur, (3.27), est la somme des carrés des résidus du modèle non contraint (3.19). Ces résidus sont ce qui reste après que  $\mathbf{u}$  ait été projeté en dehors de tout ce qui se situe en  $\mathcal{S}(\mathbf{X})$ . Par contraste, la forme quadratique qui apparaît au numérateur, (3.23), est la somme des réductions des résidus au carré du modèle contraint réalisée en ajoutant  $\mathbf{X}_2$  à la régression. Ces réductions doivent se situer en  $\mathcal{S}(\mathbf{X})$ . Il en résulte que la variable aléatoire qui apparaît au numérateur doit se situer dans un sous-espace orthogonal à celui dans lequel se situe la variable aléatoire qui apparaît au dénominateur. Cela se voit clairement dans la partie (a) de la Figure 1.7, dans le cas précis où  $r = 1$  et  $k = 2$ .

Nous avons désormais vérifié que le Fisher (3.17) suit réellement la distribution de Fisher à  $r$  et  $n - k$  degrés de liberté, pour le cas des modèles linéaires soumis à des contraintes linéaires et des aléas normalement distribués. Un corollaire simple de ce résultat est que pour le même cas le Student (3.12) ordinaire suit la distribution de Student à  $n - k$  degrés de liberté. Pour s'en rendre compte, supposons que  $\mathbf{X}_2$  est composé d'une colonne unique, que nous appellerons  $\mathbf{x}_2$ . Alors avec le Théorème FWL, l'estimation de  $\beta_2$  à partir de (3.19) est identique à l'estimation à partir de la régression

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{x}_2 \beta_2 + \text{résidus},$$



qui est égale à

$$(\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{-1} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}.$$

L'écart type estimé de  $\hat{\beta}_2$  est  $s(\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{-1/2}$ . De sorte que le Student pour  $\beta_2 = 0$  est

$$\frac{\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}}{s(\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{1/2}}.$$

Le carré de ce Student est

$$\frac{1}{s^2} \mathbf{y}^\top \mathbf{M}_1 \mathbf{x}_2 (\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{-1} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}, \quad (3.28)$$

et c'est tout simplement le Fisher (3.17) pour le cas particulier où  $r = 1$ . Puisque la racine carrée d'une variable aléatoire qui suit une Fisher  $F(1, n-k)$  suit une distribution de Student  $t(n-k)$  (consulter l'Annexe A), nous avons prouvé le corollaire que nous avons mis en évidence.

La géométrie des Students et Fisher est intéressante. Le carré du Student que nous avons observé est donné par (3.28). Elle dépend de  $s$ , l'estimation OLS de  $\sigma$ , qui est donnée par

$$s \equiv \left( \frac{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}}{n-k} \right)^{1/2}. \quad (3.29)$$

Du Théorème FWL nous savons que

$$\mathbf{y}^\top \mathbf{M}_X \mathbf{y} = \mathbf{y}^\top \mathbf{M}_1 \mathbf{M}_{M_1 \mathbf{x}_2} \mathbf{M}_1 \mathbf{y}, \quad (3.30)$$

où  $\mathbf{M}_{M_1 \mathbf{x}_2}$  désigne la matrice qui projette en dehors de  $\mathcal{S}(\mathbf{M}_1 \mathbf{x}_2)$ . En d'autres termes, cela signifierait simplement que la somme des résidus au carré à partir de la régression de  $\mathbf{y}$  sur  $\mathbf{X}$  est la même que la somme des résidus au carré à partir de la régression de  $\mathbf{M}_1 \mathbf{y}$  sur  $\mathbf{M}_1 \mathbf{x}_2$ . En utilisant (3.29) et (3.30), il nous est possible de réécrire le Student (3.28) au carré comme

$$\begin{aligned} & (n-k) \frac{\mathbf{y}^\top \mathbf{M}_1 \mathbf{x}_2 (\mathbf{x}_2^\top \mathbf{M}_1 \mathbf{x}_2)^{-1} \mathbf{x}_2^\top \mathbf{M}_1 \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_1 \mathbf{M}_{M_1 \mathbf{x}_2} \mathbf{M}_1 \mathbf{y}} \\ &= (n-k) \frac{\mathbf{y}^\top \mathbf{M}_1 \mathbf{P}_{M_1 \mathbf{x}_2} \mathbf{M}_1 \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_1 \mathbf{M}_{M_1 \mathbf{x}_2} \mathbf{M}_1 \mathbf{y}}, \end{aligned} \quad (3.31)$$

de sorte que le Student lui-même est

$$\text{signe}(\mathbf{y}^\top \mathbf{M}_1 \mathbf{x}_2) (n-k)^{1/2} \frac{\|\mathbf{P}_{M_1 \mathbf{x}_2} \mathbf{M}_1 \mathbf{y}\|}{\|\mathbf{M}_{M_1 \mathbf{x}_2} \mathbf{M}_1 \mathbf{y}\|}.$$

Nous voyons donc que les Students ont une interprétation géométrique simple. Si nous disons que  $\mathbf{y}$  sur la Figure 1.3 représente  $\mathbf{M}_1 \mathbf{y}$ , et  $\mathcal{S}(\mathbf{X})$  représente

$\mathcal{S}(\mathbf{M}_1\mathbf{x}_2)$ , alors il est clair que le Student de  $\beta_2$  est simplement  $(n - k)^{1/2}$  fois la cotangente de l'angle  $\phi$  (consulter l'Annexe A). Lorsque l'angle est nul, et donc  $\mathbf{M}_1\mathbf{x}_2$  explique entièrement  $\mathbf{M}_1\mathbf{y}$ , le Student est égal à, soit plus l'infini, soit moins l'infini, son signe dépendant du signe de  $\hat{\beta}_2$ . Au contraire, lorsque l'angle est de  $90^\circ$ , et donc  $\mathbf{M}_1\mathbf{x}_2$  a un pouvoir explicatif nul sur  $\mathbf{M}_1\mathbf{y}$ , le Student est nul. Pour tout angle  $\phi$  compris entre zéro et un, exclus, l'amplitude du Student sera proportionnelle à  $(n - k)^{1/2}$ , de sorte que si  $\mathbf{M}_1\mathbf{x}_2$  a un quelconque pouvoir explicatif sur  $\mathbf{M}_1\mathbf{y}$ , nous nous attendrions à ce que le Student nous permette de rejeter l'hypothèse nulle avec la probabilité unitaire lorsque l'échantillon est de taille suffisamment importante.

Les résultats précédents s'appliquent bien entendu également aux statistiques  $F$  moyennant quelques modifications légères. Si  $\mathbf{X}_2$  possède  $r$  colonnes au lieu d'une seule, le Fisher pour tester  $\beta_2 = \mathbf{0}$  peut s'écrire sous une forme équivalente à (3.31), c'est-à-dire

$$\begin{aligned} & \frac{n - k}{r} \times \frac{\mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_1 \mathbf{M}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y}} \\ &= \frac{n - k}{r} \times \frac{\|\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y}\|^2}{\|\mathbf{M}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1 \mathbf{y}\|^2}. \end{aligned} \quad (3.32)$$

Par conséquent, le Fisher est égal à  $(n - k)/r$  fois le carré de la cotangente de l'angle  $\phi$  formé par les vecteurs  $\mathbf{M}_1\mathbf{y}$  et  $\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{M}_1\mathbf{y}$ ; observons à nouveau la Figure 1.3 en gardant à l'esprit que  $\mathbf{M}_1\mathbf{y}$  remplace désormais  $\mathbf{y}$  et  $\mathcal{S}(\mathbf{M}_1\mathbf{X}_2)$  remplace  $\mathcal{S}(\mathbf{X})$ . Comme nous l'avons vu au cours du Chapitre 1, le carré du cosinus de  $\phi$  est le  $R^2$  de ce qui est ici la régression de  $\mathbf{M}_1\mathbf{y}$  sur  $\mathbf{M}_1\mathbf{X}_2$ . Ceci suggère un lien étroit entre les Fisher et les  $R^2$ . Ce lien existe en effet.

Un usage familier du test en  $F$  est de tester l'hypothèse nulle selon laquelle tous les coefficients sauf le terme constant sont nuls. C'est un test de  $\beta_2 = \mathbf{0}$  dans le modèle de régression

$$\mathbf{y} = \beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u}, \quad (3.33)$$

où  $\mathbf{X}_2$  est désormais  $n \times (k - 1)$ . Avec le Théorème FWL, la SSR de cette régression est identique à celle de la régression

$$\mathbf{M}_l \mathbf{y} = \mathbf{M}_l \mathbf{X}_2 \beta_2 + \text{résidus},$$

qui est

$$\mathbf{y}^\top \mathbf{M}_l \mathbf{y} - \mathbf{y}^\top \mathbf{M}_l \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_l \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_l \mathbf{y}.$$

Ainsi la différence entre la somme des résidus contraints et la somme des résidus non contraints est

$$\mathbf{y}^\top \mathbf{M}_l \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_l \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_l \mathbf{y} = \|\mathbf{P}_{\mathbf{M}_l \mathbf{X}_2} \mathbf{M}_l \mathbf{y}\|^2,$$

et le Fisher pour l'hypothèse  $\beta_2 = \mathbf{0}$  est

$$\frac{n-k}{k-1} \times \frac{\|\mathbf{P}_{M_{\ell}X_2}\mathbf{M}_{\ell}\mathbf{y}\|^2}{\|\mathbf{M}_{M_{\ell}X_2}\mathbf{M}_{\ell}\mathbf{y}\|^2}, \quad (3.34)$$

qui est un cas particulier de (3.32).

Parce que (3.33) contient effectivement un terme constant, le  $R^2$  centré pour cette régression est

$$R_c^2 = \frac{\|\mathbf{P}_{M_{\ell}X_2}\mathbf{M}_{\ell}\mathbf{y}\|^2}{\|\mathbf{M}_{\ell}\mathbf{y}\|^2}.$$

Evidemment cela correspond, comme d'habitude, au carré du cosinus de l'angle formé par  $\mathbf{P}_{M_{\ell}X_2}\mathbf{M}_{\ell}\mathbf{y}$  et  $\mathbf{M}_{\ell}\mathbf{y}$ , alors que le Fisher (3.34) est égal à  $(n-k)/(k-1)$  fois le carré de la cotangente du même angle. Il est ainsi envisageable d'exprimer ce Fisher comme

$$\frac{n-k}{k-1} \times \frac{R_c^2}{1-R_c^2}.$$

Ce résultat peut être mis en évidence si l'on considère que  $\cot^2\phi = \cos^2\phi/(1 - \cos^2\phi)$  ou par l'usage des définitions algébriques simples de  $F$  et de  $R_c^2$ . Notons toutefois qu'il faudrait normalement éviter de calculer un Fisher de cette façon, à moins qu'à la fois  $R_c^2$  et  $1 - R_c^2$  ne soient connus avec autant de décimales qu'il est nécessaire pour construire le Fisher avec précision. Plusieurs progiciels de régression rapportent le  $R^2$  sous le format d'un nombre à trois ou quatre décimales, et  $1 - R^2$  peut être précis à une seule ou à deux décimales.

Il est possible d'exprimer le  $R^2$  comme

$$R_c^2 = 1 - \frac{\|\mathbf{M}_{M_{\ell}X_2}\mathbf{M}_{\ell}\mathbf{y}\|^2}{\|\mathbf{M}_{\ell}\mathbf{y}\|^2} = 1 - \frac{\|\mathbf{M}_X\mathbf{y}\|^2/n}{\|\mathbf{M}_{\ell}\mathbf{y}\|^2/n}.$$

Le numérateur du deuxième terme du membre de droite est ici  $\hat{\sigma}^2$ , dont nous savons que c'est une estimation biaisée de  $\sigma^2$ . Le dénominateur est aussi une estimation biaisée de la variance de  $y_t$  autour de son espérance conditionnelle. Il paraît naturel de remplacer ces estimations biaisées par des estimations non biaisées. Cette opération nous permet de disposer du  $\bar{R}^2$ , le  **$R^2$  ajusté**, dont la définition est

$$\bar{R}^2 = 1 - \frac{\|\mathbf{M}_X\mathbf{y}\|^2/(n-k)}{\|\mathbf{M}_{\ell}\mathbf{y}\|^2/(n-1)} = 1 - \frac{n-1}{n-k} \frac{\|\mathbf{M}_X\mathbf{y}\|^2}{\|\mathbf{M}_{\ell}\mathbf{y}\|^2}.$$

Le  $\bar{R}^2$  non centré est extrêmement rare, aussi avons-nous omis d'indicer le  $\bar{R}^2$  par un "c".

La quantité  $\bar{R}^2$  est disponible dans presque tous les progiciels de régression. Ce n'est pas cependant une estimation d'un quelconque paramètre du modèle (puisque pour la grande majorité des modèles, la variance de  $y_t$  autour de sa propre espérance non conditionnelle dépendra de la distribution des variables du membre de droite), et ne sera pas particulièrement utile en pratique. L'usage grandissant du  $\bar{R}^2$  date des balbutiements de l'économétrie, lorsque les échantillons étaient de taille modeste et les chercheurs souvent impressionnés par des modèles qui s'ajustaient bien, dans le sens où la valeur du  $R_c^2$  était importante. Les observateurs ont rapidement appris qu'en ajoutant des régresseurs dans une régression linéaire, on augmente toujours le  $R^2$  ordinaire (centré ou non), et plus particulièrement lorsque la taille de l'échantillon est faible. Cela poussa quelques chercheurs à estimer plus rigoureusement les modèles intégrant un grand nombre de paramètres. On préconise l'usage du  $\bar{R}^2$  plutôt que le  $R_c^2$  pour traiter ce type de problème, parce que l'addition de nouveaux régresseurs ne fera augmenter le  $\bar{R}^2$  qu'à condition que la réduction proportionnelle de la SSR soit plus forte que la réduction proportionnelle de  $n - k$ .

Comme nous l'avons noté au tout début de cette section, tous les résultats exacts sur la distribution des Students et Fisher (ou plus généralement sur les distributions des estimateurs OLS avec des échantillons de taille finie) requièrent que les aléas soient normaux et que les régresseurs soient fixés en échantillons répétés, ou qu'ils puissent être traités comme tels. La dernière possibilité nécessite quelques commentaires. La raison pour laquelle il est pratique de supposer que les régresseurs sont fixes est que nous voulons être capables de manipuler les expressions matricielles qui dépendent des régresseurs comme des expressions invariables en vue d'opérer des calculs d'espérances. Aussi grâce à cette hypothèse, il nous est possible de soutenir que, par exemple

$$E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{u}) = \mathbf{0}. \quad (3.35)$$

L'hypothèse de régresseurs fixes nous permet de faire cette manipulation, mais elle s'avère malheureusement trop forte dans la plupart des applications économétriques.

Une hypothèse plus simple, mais qui a le même effet, est celle selon laquelle toutes les espérances que nous calculons sont conditionnelles à la matrice  $\mathbf{X}$ , de sorte que, par exemple,  $E(\mathbf{u})$  dans (3.35) doit s'interpréter comme  $E(\mathbf{u} | \mathbf{X})$ . Puisque nous conditionnons par rapport à  $\mathbf{X}$ , nous la traitons véritablement comme fixe. Cependant, à moins que la matrice  $\mathbf{X}$  ne soit totalement indépendante des aléas  $\mathbf{u}$ , les Students et Fisher ne seront pas distribués suivant les lois dont ils portent le nom. Le problème survient dans le cas des modèles dynamiques, dans lesquels les aléas ne peuvent pas être indépendants des valeurs retardées des variables dépendantes.

### 3.6 TESTS D'HYPOTHÈSES DANS LE CAS NON LINÉAIRE

Il y a au moins trois manières différentes de dériver les statistiques de test pour des hypothèses concernant les paramètres des modèles de régression non linéaire. Elles consistent à employer le **principe de Wald**, le **principe du multiplicateur de Lagrange**, et le **principe du rapport de vraisemblance**. Elles offrent ce que l'on regroupe souvent sous la dénomination de statistiques de test “classiques”. Dans cette section, nous présentons ces trois principes et montrons comment ils produisent des statistiques de test pour les hypothèses relatives à  $\beta$  dans les modèles de régression non linéaire (et implicitement dans les modèles de régression linéaire également, puisque les modèles linéaires ne sont que des cas particuliers des modèles non linéaires). Les trois principes peuvent avoir un emploi très étendu, et réapparaîtront dans des contextes variés au cours de l'ouvrage.<sup>7</sup> Un traitement formel de ces tests dans le cadre des moindres carrés sera offert au Chapitre 5. Ils seront ensuite introduits dans le cadre de l'estimation par maximum de vraisemblance au Chapitre 8, et une discussion plus détaillée dans ce même contexte suivra au Chapitre 13. Les références sérieuses que l'on peut donner sont Engle (1984) et Godfrey (1988), et une discussion introductive se trouvera chez Buse (1982).

Le principe de Wald, dû à Wald (1943), consiste à élaborer une statistique de test sur les estimations des paramètres non contraints et une estimation de la matrice de covariance non contrainte. Si l'hypothèse n'implique qu'une seule contrainte, disons  $\beta_i = \beta_i^*$ , alors il est possible de calculer la **statistique pseudo- $t$**

$$\frac{\hat{\beta}_i - \beta_i^*}{\hat{S}(\hat{\beta}_i)}. \quad (3.36)$$

On se réfère à cette statistique en tant que statistique “pseudo- $t$ ” parce que sa distribution ne sera pas véritablement la Student à  $n - k$  degrés de liberté avec des échantillons finis, lorsque  $x_t(\beta)$  est non linéaire dans ses paramètres, lorsque  $x_t(\beta)$  dépend des valeurs retardées  $y_t$ , ou lorsque les aléas  $u_t$  ne sont pas normalement distribués. Cependant, elle sera distribuée asymptotiquement suivant une  $N(0, 1)$  moyennant quelques hypothèses légères (voir Chapitre 5), et sa distribution dans le cas d'un échantillon fini sera fréquemment approximée avec satisfaction par la Student à  $n - k$  degrés de liberté.

Dans le contexte plus général où il y a  $r$  contraintes au lieu d'une seule, les tests de Wald emploieraient le fait que si  $v$  est un vecteur aléatoire à  $r$  com-

<sup>7</sup> Nous employons les termes de “principe du multiplicateur de Lagrange”, “principe de Wald”, et “principe du rapport de vraisemblance” plutôt que les termes “tests du multiplicateur de Lagrange”, “tests de Wald”, et “tests du rapport de vraisemblance” parce que de nombreux auteurs font usage de ces derniers termes pour faire référence à des tests basés sur les modèles estimés par la méthode du maximum de vraisemblance (voir Chapitre 8 et 13). Nous croyons que cet usage ambigu se perdra lorsque la polyvalence de ces trois principes sera largement reconnue.

posantes distribuées normalement avec une espérance nulle et une matrice de covariance  $\mathbf{A}$ , la forme quadratique

$$\mathbf{v}^\top \mathbf{A}^{-1} \mathbf{v} \quad (3.37)$$

doit suivre une distribution du  $\chi^2(r)$ . Ce résultat est démontré dans l'Annexe B et nous l'avons déjà utilisé dans les Sections 3.3 et 3.5.

Pour élaborer un test de Wald asymptotique, il suffit ensuite de trouver un vecteur de variables aléatoires qui, sous l'hypothèse nulle, soit asymptotiquement normalement distribué avec une espérance nulle et une matrice de covariance que l'on peut estimer. Par exemple, supposons que  $\boldsymbol{\beta}$  soit soumis à  $r(\leq k)$  contraintes linéairement indépendantes,

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \quad (3.38)$$

où  $\mathbf{R}$  est une matrice de dimension  $r \times k$  de rang  $r$ , et  $\mathbf{r}$  est un vecteur à  $r$  composantes. Nous n'avons supposé les contraintes linéaires que par souci de simplicité, et non pas parce que les tests de Wald ne peuvent pas intégrer des contraintes non linéaires. Toutefois, parce que les tests de Wald ne sont pas invariants à une reparamétrisation non linéaire du modèle ou des contraintes, il faut prendre quelques précautions lorsque l'on teste des contraintes non linéaires à l'aide de tels tests. Nous verrons tout ceci au cours du Chapitre 13; consulter Gregory et Veall (1985), Lafontaine et White (1986), et Phillips et Park (1988). Ainsi, il paraît approprié de ne se concentrer que sur le cas linéaire pour l'instant.

Supposons que l'on évalue le vecteur  $\mathbf{R}\boldsymbol{\beta} - \mathbf{r}$  pour les estimations non contraintes  $\hat{\boldsymbol{\beta}}$  pour obtenir un vecteur aléatoire à  $r$  composantes

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}. \quad (3.39)$$

Comme nous le démontrerons au Chapitre 5, si les données ont été générées réellement par le modèle que l'on teste, le vecteur d'estimations  $\hat{\boldsymbol{\beta}}$  tendrait asymptotiquement vers le vrai vecteur de paramètres  $\boldsymbol{\beta}_0$ , et la matrice de covariance de  $\hat{\boldsymbol{\beta}}$  autour de  $\boldsymbol{\beta}_0$  pourrait être évaluée avec satisfaction par  $s^2(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1}$ . Si les contraintes (3.38) sont effectives, alors il doit être vérifié que

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} = \mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \mathbf{R}\boldsymbol{\beta}_0 - \mathbf{r} = \mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

de sorte que chaque élément de  $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$  est une combinaison linéaire des éléments de  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ . Ainsi la matrice de covariance de (3.39) doit être

$$\mathbf{V}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) = \mathbf{R}\mathbf{V}(\hat{\boldsymbol{\beta}})\mathbf{R}^\top,$$

qui peut être estimée par la matrice

$$s^2 \mathbf{R}(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \mathbf{R}^\top. \quad (3.40)$$

En combinant (3.39) et (3.40), on obtient la statistique de Wald

$$\begin{aligned} & (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top (s^2 \mathbf{R}(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \mathbf{R}^\top)^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \\ &= \frac{1}{s^2} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})^\top (\mathbf{R}(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \mathbf{R}^\top)^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}). \end{aligned} \quad (3.41)$$

Pourvu que  $\hat{\boldsymbol{\beta}}$  soit asymptotiquement normalement distribué et que  $\mathbf{V}(\hat{\boldsymbol{\beta}})$  converge asymptotiquement vers la véritable matrice de covariance, alors (3.41) sera asymptotiquement distribuée suivant la loi du  $\chi^2(r)$ . Notons que dans le cas simple où  $\mathbf{R}$  est un vecteur colonne composé de zéros et d'un un à la  $i$ ème position, et  $\mathbf{r}$  est égal à  $\beta_i^*$ , de sorte que l'on teste l'hypothèse  $\beta_i = \beta_i^*$ , le carré de la statistique pseudo- $t$  (3.36) est précisément la statistique de Wald (3.41).

La deuxième approche pour calculer des statistiques de test consiste à estimer le modèle soumis aux contraintes que l'on doit tester. On appelle souvent cette approche principe du multiplicateur de Lagrange (ou LM), parce qu'une façon d'obtenir les estimations moindres carrés est d'établir un Lagrangien qui est simultanément minimisé par rapport aux paramètres et maximisé par rapport aux multiplicateurs de Lagrange. Lorsque les contraintes sont vraies, on s'attend à ce que les multiplicateurs de Lagrange estimés aient une espérance nulle (du moins asymptotiquement); l'idée générale des tests LM est de savoir si c'est véritablement le cas.

Pour estimer un modèle tel que (3.01) soumis aux contraintes (3.38), il est possible d'élaborer le Lagrangien

$$\frac{1}{2}(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}))^\top (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})) + (\mathbf{R}\boldsymbol{\beta} - \mathbf{r})^\top \boldsymbol{\lambda}, \quad (3.42)$$

où  $\boldsymbol{\lambda}$  est un vecteur composé de  $r$  multiplicateurs de Lagrange, et où la fonction  $SSR(\boldsymbol{\beta})$  a été multipliée par .5 pour simplifier l'expression algébrique. Les conditions du premier ordre que l'on obtient en dérivant (3.42) par rapport à  $\boldsymbol{\beta}$  et  $\boldsymbol{\lambda}$ , et en annulant ces dérivées partielles, sont

$$-\mathbf{X}^\top(\tilde{\boldsymbol{\beta}})(\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})) + \mathbf{R}^\top \tilde{\boldsymbol{\lambda}} = \mathbf{0} \quad (3.43)$$

$$\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r} = \mathbf{0}, \quad (3.44)$$

où  $\tilde{\boldsymbol{\beta}}$  désigne les estimations contraintes et  $\tilde{\boldsymbol{\lambda}}$  désigne les multiplicateurs de Lagrange estimés. A partir de (3.43) nous voyons que

$$\mathbf{R}^\top \tilde{\boldsymbol{\lambda}} = \tilde{\mathbf{X}}^\top (\mathbf{y} - \tilde{\mathbf{x}}), \quad (3.45)$$

où selon notre notation habituelle,  $\tilde{\mathbf{x}}$  et  $\tilde{\mathbf{X}}$  désignent  $\mathbf{x}(\tilde{\boldsymbol{\beta}})$  et  $\mathbf{X}(\tilde{\boldsymbol{\beta}})$ . Le membre de droite de (3.45) est un vecteur composé de  $k$  dérivées de  $-\frac{1}{2}SSR(\boldsymbol{\beta})$  par rapport aux éléments de  $\boldsymbol{\beta}$ , évaluées en  $\tilde{\boldsymbol{\beta}}$ . Ce vecteur est souvent dénommé **vecteur du score**. Puisque  $\mathbf{y} - \tilde{\mathbf{x}}$  correspond au vecteur de résidus, qui devrait

converger asymptotiquement sous  $H_0$  vers le vecteur des aléas  $\mathbf{u}$ , il semble plausible que la matrice de covariance du vecteur du score soit

$$\sigma_0^2 \mathbf{X}^\top(\beta_0) \mathbf{X}(\beta_0). \quad (3.46)$$

Moyennant quelques propriétés asymptotiques, c'est en effet le cas, et une version plus rigoureuse de ce résultat sera présentée au cours du Chapitre 5.

Le moyen le plus évident d'évaluer (3.46) est de faire usage de  $\tilde{s}^2 \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ , où  $\tilde{s}^2$  est égale à  $SSR(\tilde{\beta})/(n - k + r)$ . En combinant cette estimation avec les deux membres de l'égalité (3.45), nous pouvons construire deux statistiques de test qui diffèrent à première vue, mais qui sont numériquement identiques. La première d'entre elles est

$$\tilde{\lambda}^\top \mathbf{R}(\tilde{s}^2 \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \mathbf{R}^\top \tilde{\lambda} = \frac{1}{\tilde{s}^2} \tilde{\lambda}^\top \mathbf{R}(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \mathbf{R}^\top \tilde{\lambda}. \quad (3.47)$$

Sous cette forme, on a clairement à faire à une statistique du multiplicateur de Lagrange. Puisque  $\tilde{\lambda}$  est un vecteur à  $r$  composantes, il ne devrait pas être surprenant que cette statistique suive une distribution de  $\chi^2(r)$ . Pour démontrer cette proposition, nous utilisons pour l'essentiel les mêmes arguments que ceux employés pour les tests de Wald, puisque (3.47) est une forme quadratique similaire à (3.37). Evidemment, ce résultat dépend du vecteur  $\tilde{\lambda}$  qui doit être asymptotiquement normal, ce que nous démontrerons au Chapitre 5.

La seconde statistique de test dont nous avons précisé l'identité numérique avec la première, est obtenue en remplaçant  $\tilde{\mathbf{X}}^\top(\mathbf{y} - \tilde{\mathbf{x}})$  par  $\tilde{\lambda}^\top \mathbf{R}$  dans (3.47). Le résultat, qui correspond à la **forme du score** de la statistique LM, est

$$\frac{1}{\tilde{s}^2} (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}} (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top (\mathbf{y} - \tilde{\mathbf{x}}) = \frac{1}{\tilde{s}^2} (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{P}}_X (\mathbf{y} - \tilde{\mathbf{x}}), \quad (3.48)$$

où  $\tilde{\mathbf{P}}_X \equiv \tilde{\mathbf{X}} (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top$ . Clairement, cette expression est tout simplement la somme des carrés expliqués de la **régression linéaire artificielle**

$$\frac{1}{\tilde{s}} (\mathbf{y} - \tilde{\mathbf{x}}) = \tilde{\mathbf{X}} \mathbf{b} + \text{résidus}, \quad (3.49)$$

pour laquelle les résidus  $\mathbf{y} - \tilde{\mathbf{x}}$ , que l'on a préalablement divisés par  $\tilde{s}$ , l'écart type estimé de la régression contrainte, sont régressés sur la matrice  $\tilde{\mathbf{X}}$ . La régression (3.49) est un exemple de la **régression de Gauss-Newton**, ou **GNR**, dont nous ferons un exposé très détaillé dans le Chapitre 6. Il est clair que la forme du score de la statistique LM peut se calculer aisément une fois que l'on a obtenu les estimations contraintes  $\tilde{\beta}$ , avec l'aide explicite ou pas du Lagrangien (3.42). Les tests LM peuvent presque toujours se calculer à l'aide des régressions linéaires artificielles, comme nous le verrons au Chapitre 6 pour les modèles de régression non linéaires, et au Chapitres 13, 14 et 15 pour les modèles estimés par maximum de vraisemblance.



La troisième et dernière approche que l'on peut choisir pour tester les hypothèses concernant  $\beta$  est d'estimer à la fois le modèle soumis aux contraintes et le modèle non contraint, et ainsi d'obtenir deux valeurs de  $SSR(\beta)$ , que nous désignerons par  $SSR(\tilde{\beta})$  et  $SSR(\hat{\beta})$ . Une statistique pseudo- $F$  peut alors être calculée comme

$$\frac{(SSR(\tilde{\beta}) - SSR(\hat{\beta}))/r}{SSR(\hat{\beta})/(n - k)} = \frac{SSR(\tilde{\beta}) - SSR(\hat{\beta})}{rs^2}. \quad (3.50)$$

Nous avons déjà vu que cette statistique de test suit exactement une distribution de  $F(r, n - k)$  lorsqu'elle sert à tester les contraintes linéaires sur les modèles linéaires où les aléas sont normaux. Nous aurons l'occasion de démontrer au Chapitre 5 qu'elle suit en général la distribution de  $F(r, n - k)$  (et également que multipliée par  $r$ , elle suit une distribution du  $\chi^2(r)$ ), asymptotiquement. Comme nous l'observerons au Chapitre 6, des statistiques pseudo- $F$  asymptotiquement correctes peuvent aussi être calculées à partir de régressions artificielles similaires à (3.49), mais pour lesquelles la régressande n'a pas été nécessairement divisée par  $\hat{s}$ . Nous remplaçons tout simplement  $SSR(\hat{\beta})$  dans (3.50) par la somme des résidus au carré de la régression artificielle.

L'idée générale de la statistique de test (3.50) est d'observer la différence entre les valeurs de la fonction objectif  $SSR(\beta)$  en les estimations contraintes et non contraintes. Le dénominateur a pour fonction de normaliser les résultats. Plus tard, dans cet ouvrage, nous verrons que l'on peut trouver des tests d'hypothèses basés sur les logarithmes des fonctions de vraisemblance, lorsque nous discuterons de l'estimation par maximum de vraisemblance. Parce que la différence de deux logarithmes est le logarithme d'un rapport (dans ce cas le rapport de la valeur de la fonction de vraisemblance contrainte par la valeur de la fonction de vraisemblance non contrainte), ceux-ci sont connus sous le nom de **tests du rapport de vraisemblance**, ou **tests LR**. Ce serait faire un usage abusif de la terminologie que d'appeler (3.50) un test LR, mais il est certainement raisonnable de prétendre que (3.50) est fondé sur le principe du rapport de vraisemblance si ce dernier est grossièrement défini comme désignant un test basé sur la différence de deux valeurs de la fonction objectif calculée pour les estimations contraintes et non contraintes.

Une caractéristique notable des principes de Wald, de LM et de LR est que des tests qui s'appuient sur ces principes sont **asymptotiquement équivalents**. De façon intuitive, cela signifie que lorsque la taille de l'échantillon devient suffisamment importante et que les hypothèses testées sont soit correctes, soit presque correctes (nous définirons plus précisément ce que nous entendons par presque correctes, aux Chapitres 12 et 13), alors les statistiques de test de la même hypothèse nulle basées sur n'importe lequel des trois principes devraient donner exactement les mêmes résultats. Bien entendu, avec des échantillons finis et des modèles qui pourraient ne pas être exacts, les trois principes peuvent souvent donner des tests qui conduisent à

des résultats différents. De fait, le choix du test dépend souvent du test dont la distribution en échantillon fini est le mieux approchée par la distribution en échantillon infini.

Une démonstration de l'équivalence asymptotique des tests LM, LR et Wald dans un contexte de modèles de régression non linéaire dépasse la portée de ce chapitre. Toutefois, dans le cas particulier des modèles de régression linéaire, une telle démonstration est assez aisée. Par souci de simplicité, nous supposons que l'on est confronté à la situation décrite dans la Section 3.5, puisque nous n'essayons pas d'extraire des résultats exacts. L'hypothèse nulle est  $\beta_2 = \mathbf{0}$  dans la régression

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}).$$

Comme nous l'avons vu dans la Section 3.5, l'utilisation du principe du rapport de vraisemblance dans ce problème de test entraîne le Fisher

$$\frac{\|\mathbf{P}_{M_1X_2}\mathbf{y}\|^2}{rs^2}, \quad (3.51)$$

qui, multiplié par  $r$  suivrait asymptotiquement la distribution du  $\chi^2(r)$ . Evidemment, il n'y a rien dans le principe de LR qui nous oblige à choisir un test en  $F$  plutôt qu'un test en  $\chi^2$ , et rien dans les principes de Wald et du LM ne nous oblige à employer les tests en  $\chi^2$  plutôt que les tests en  $F$ . Le choix entre les configurations  $F$  et  $\chi^2$  devrait normalement se fonder sur la taille de l'échantillon, ce qui favorise la configuration en  $F$  lorsqu'elle est finie.

Maintenant examinons ce qu'il advient lorsque nous appliquons les principes de Wald et du LM. La formule générale pour un test de Wald sur un modèle de régression non linéaire est donnée par (3.41). Dans ce cas  $\mathbf{R}\hat{\beta} - \mathbf{r}$  correspond à  $\hat{\beta}_2$ . On peut faire usage du Théorème FWL pour obtenir une expression de ce dernier vecteur, et ainsi avoir

$$\mathbf{R}\hat{\beta} - \mathbf{r} = \hat{\beta}_2 = (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}.$$

La matrice  $\mathbf{R}$  dans ce cas est  $[\mathbf{0}_{k-r} \quad \mathbf{I}_r]$ . De sorte que  $\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{R}^\top$ , que nous avons besoin d'évaluer pour élaborer une statistique de Wald, est simplement le bloc de dimension  $r \times r$  situé en bas à droite de la matrice  $(\mathbf{X}^\top \mathbf{X})^{-1}$ . Il est aisé d'arriver à ce résultat en faisant usage du Théorème FWL ou des formules sur les inversions des matrices partitionnées que l'on a rappelées dans l'Annexe A; c'est tout simplement  $(\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1}$ . Par conséquent, la statistique de Wald (3.41) est

$$\begin{aligned} & \frac{1}{s^2} \mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} \\ &= \frac{1}{s^2} \mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} = \frac{1}{s^2} \|\mathbf{P}_{M_1X_2}\mathbf{y}\|^2, \end{aligned} \quad (3.52)$$

qui est égal à  $r$  fois (3.51). Ainsi, dans ce cas, les principes de Wald et du LR entraînent pour l'essentiel les mêmes statistiques de test. La seule différence provient du fait que nous choisissons d'écrire la statistique de Wald (3.52) dans la configuration du  $\chi^2$  lorsque la statistique LR (3.51) s'écrit dans une configuration en  $F$ .

Qu'en est-il du principe du LM? Nous avons vu que la statistique LM (3.18) correspond à la somme des carrés expliqués de la régression artificielle (3.49), qui dans ce cas est

$$\frac{1}{\tilde{s}} \mathbf{M}_1 \mathbf{y} = \mathbf{X}_1 \mathbf{b}_1 + \mathbf{X}_2 \mathbf{b}_2 + \text{résidus.} \quad (3.53)$$

Puisque la régressande est ici orthogonale à  $\mathbf{X}_1$ , la somme des carrés expliqués de (3.53) doit, par l'intermédiaire du Théorème FWL, être égale à la somme des carrés expliqués de la régression

$$\frac{1}{\tilde{s}} \mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \mathbf{X}_2 \mathbf{b}_2 + \text{résidus,}$$

qui est

$$\frac{1}{\tilde{s}^2} \mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} = \frac{1}{\tilde{s}^2} \|\mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y}\|^2.$$

Ainsi, pour les contraintes linéaires s'appliquant à des modèles de régression linéaire, la seule différence entre les statistiques LM et les statistiques de Wald et LR est que la première estime  $\sigma^2$  par  $\tilde{s}^2$  alors que les dernières utilisent  $s^2$ . Si  $\sigma^2$  était connue, les trois statistiques de test seraient identiques. Si l'hypothèse nulle était vérifiée, à la fois  $\tilde{s}^2$  et  $s^2$  tendraient vers  $\sigma^2$  à mesure que la taille de l'échantillon augmente. Ainsi, on considère les trois statistiques de test comme équivalentes asymptotiquement. Même lorsque les échantillons sont de taille finie, on s'attend à ce que  $\tilde{s}^2$  et  $s^2$  soient très proches l'une de l'autre lorsque  $H_0$  est vraie, à moins que la taille de l'échantillon ne soit extrêmement petite. Normalement les différences substantielles entre les trois statistiques de test sont très peu probables si l'hypothèse nulle était en fait vérifiée. Bien entendu, si l'hypothèse nulle était fausse,  $\tilde{s}^2$  et  $s^2$  pourraient diverger sensiblement, et par conséquent la statistique LM serait, dans ce cas, relativement différente des deux autres statistiques de test.

### 3.7 CONTRAINTES ET ESTIMATEURS DE TESTS PRÉLIMINAIRES

Dans les trois sections qui ont précédé, nous avons discuté des tests d'hypothèses en détail, mais nous n'avons rien dit sur les motivations profondes qui font que l'on impose et que l'on teste des restrictions. Dans la plupart des cas, ce n'est pas la théorie économique qui impose des contraintes, mais plutôt le chercheur qui les impose dans le but d'obtenir un modèle contraint plus facile

à estimer et qui produira des estimations plus efficaces que le modèle non contraint. Les tests de ce genre de contraintes sont les tests DWH (Chapitre 7), les tests d'autocorrélation (Chapitre 10), les tests de contrainte du facteur commun (Chapitre 10), les tests de changement de régime (Chapitre 11), et les tests sur la longueur des retards (Chapitre 19). Dans ces situations, et dans d'autres encore, les contraintes sont testées dans le but d'arbitrer entre les modèles à utiliser pour l'inférence sur les paramètres qui nous intéressent, mais aussi pour écarter des modèles qui s'avèrent incompatibles avec les données. Toutefois, comme l'estimation et les tests se fondent sur les mêmes ensembles de données, les propriétés des estimations ultimes peuvent être très difficiles à analyser. C'est le problème des **tests préliminaires**.

Par simplification, nous ne considérerons dans cette section que le cas des modèles de régression linéaire avec des régresseurs fixés, et dont certains paramètres sont soumis à des contraintes de nullité. Le modèle contraint sera (3.18), pour lequel  $\mathbf{y}$  est régressée sur la matrice  $\mathbf{X}_1$ , de dimension  $n \times (k-r)$ , et le modèle non contraint sera (3.19), dans lequel  $\mathbf{y}$  est régressée sur  $\mathbf{X}_1$  et sur  $\mathbf{X}_2$ , qui est de dimension  $n \times r$ . Les estimations OLS des paramètres du modèle contraint sont

$$\tilde{\beta}_1 = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{y}.$$

Les estimations OLS de ces mêmes paramètres à partir du modèle non contraint s'obtiennent facilement par l'usage du Théorème FWL. Elles sont

$$\hat{\beta}_1 = (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{y},$$

où  $\mathbf{M}_2$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}_2)$ .

Il est naturel de se demander quels sont les avantages de chaque estimation par rapport à l'autre. Si les données sont effectivement générées par le DGP (3.22), qui est un cas particulier du modèle contraint, elles sont à l'évidence non biaisées. Cependant, comme nous allons le démontrer, l'estimation contrainte  $\tilde{\beta}_1$  est plus **efficace** que l'estimation non contrainte  $\hat{\beta}_1$ . Un estimateur est dit plus efficace qu'un autre si la différence entre la matrice de covariance de l'estimateur inefficace et celle de l'estimateur efficace est une matrice semi-définie positive; consulter la Section 5.5. Si  $\tilde{\beta}_1$  est plus efficace que  $\hat{\beta}_1$  selon cette définition, alors toute combinaison linéaire des éléments de  $\tilde{\beta}_1$  doit avoir une variance au plus égale à la variance correspondant à la combinaison linéaire des éléments de  $\hat{\beta}_1$ .

La démonstration de l'efficacité supérieure de  $\tilde{\beta}_1$  sur  $\hat{\beta}_1$  sous le DGP (3.22) est assez simple. La différence entre les matrices de covariance de  $\tilde{\beta}_1$  et  $\hat{\beta}_1$  est

$$\sigma_0^2 (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1} - \sigma_0^2 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}. \quad (3.54)$$

Il est aisé de montrer que cette quantité est une matrice semi-définie positive: cela peut se faire grâce à un résultat dont la démonstration apparaît dans l'Annexe A. Selon ce résultat, la différence de deux matrices symétriques

semi-définies positives est semi-définie positive si et seulement si la différence des opposées de leurs inverses est semi-définie positive. Ainsi considérons la différence

$$\mathbf{X}_1^\top \mathbf{X}_1 - \mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1 = \mathbf{X}_1^\top \mathbf{P}_2 \mathbf{X}_1, \quad (3.55)$$

où  $\mathbf{P}_2 = \mathbf{I} - \mathbf{M}_2$ . Puisque le membre de droite de (3.55) est à l'évidence une matrice semi-définie positive, ce doit être aussi le cas de (3.54).

Nous venons d'établir que l'estimateur contraint  $\tilde{\beta}_1$  est plus efficace (ou, au moins, aussi efficace) que l'estimateur non contraint  $\hat{\beta}_1$  lorsque le DGP satisfait les contraintes. Mais que se passe-t-il lorsque ce n'est pas le cas? Supposons que le DGP soit

$$\mathbf{y} = \mathbf{X}_1 \beta_{10} + \mathbf{X}_2 \beta_{20} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}), \quad (3.56)$$

avec  $\beta_{20} \neq 0$ . Il est alors facile de voir que l'estimateur contraint  $\tilde{\beta}_1$  sera, en général, biaisé. Sous ce DGP,

$$\begin{aligned} E(\tilde{\beta}_1) &= E\left((\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{y}\right) \\ &= E\left((\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top (\mathbf{X}_1 \beta_{10} + \mathbf{X}_2 \beta_{20} + \mathbf{u})\right) \\ &= \beta_{10} + (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \beta_{20}. \end{aligned} \quad (3.57)$$

A moins que  $\mathbf{X}_1^\top \mathbf{X}_2$  ne soit une matrice nulle, ou que  $\beta_{20}$  ne soit un vecteur nul,  $\tilde{\beta}_1$  sera un estimateur biaisé. L'amplitude du biais dépendra des matrices  $\mathbf{X}_1^\top \mathbf{X}_1$  et  $\mathbf{X}_1^\top \mathbf{X}_2$  et du vecteur  $\beta_{20}$ .

Des résultats très comparables à (3.57) sont valables pour tous les genres de contraintes, pas seulement pour les contraintes linéaires, et pour toutes sortes de modèles autres que les modèles de régression linéaire. Nous ne tenterons pas de traiter les modèles non linéaires pour l'instant car cela réclame un outillage technique important qui ne sera développé qu'au Chapitre 12. On trouvera des résultats analogues à (3.57) pour les modèles de régression non linéaire et pour les autres types de modèles non linéaires dans l'article de Kiefer et Skoog (1984). Le point important est que lorsque l'on impose des contraintes qui sont erronées sur certains paramètres, les estimations de tous les paramètres sont alors généralement des estimations biaisées. Et ce biais ne disparaît pas lorsque la taille de l'échantillon augmente.

Même si  $\tilde{\beta}_1$  est biaisé lorsque le DGP est (3.56), il peut être intéressant de savoir quel est son degré d'efficacité. Le concept correspondant à la matrice de covariance pour un estimateur biaisé est la **matrice d'erreur quadratique moyenne**, qui est dans ce cas

$$\begin{aligned} &E(\tilde{\beta}_1 - \beta_{10})(\tilde{\beta}_1 - \beta_{10})^\top \\ &= E(\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top (\mathbf{X}_2 \beta_{20} + \mathbf{u})(\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top (\mathbf{X}_2 \beta_{20} + \mathbf{u})^\top \\ &= \sigma_0^2 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} + (\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \mathbf{X}_1^\top \mathbf{X}_2 \beta_{20} \beta_{20}^\top \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}. \end{aligned} \quad (3.58)$$

La dernière ligne correspond à la somme de deux matrices: la matrice de covariance de  $\tilde{\beta}_1$  lorsque le DGP satisfait les contraintes, et le produit extérieur du second terme de la dernière ligne de (3.57) avec lui-même. Il est envisageable de comparer (3.58) à  $V(\hat{\beta}_1)$ , la matrice de covariance de l'estimateur non contraint  $\hat{\beta}_1$ , à condition que  $\sigma_0$  et  $\beta_{20}$  soient connus. Puisque le premier terme de (3.58) est plus petit, dans le sens matriciel du terme, que  $V(\hat{\beta}_1)$ , il est clair que si  $\beta_{20}$  est suffisamment faible (3.58) sera plus petit que  $V(\hat{\beta}_1)$ . Ainsi, il pourrait être avantageux d'utiliser  $\tilde{\beta}_1$  lorsque les contraintes sont erronées, pourvu qu'elles ne le soient pas trop.

Lors d'une application pratique on se trouve fréquemment dans une situation telle que celle décrite. On désire estimer  $\beta_1$  et on ne sait pas si, oui ou non,  $\beta_2$  est nul. Il paraît naturel de définir un nouvel estimateur,

$$\check{\beta}_1 = \begin{cases} \tilde{\beta}_1 & \text{si } F_{\beta_2=0} < c_\alpha; \\ \hat{\beta}_1 & \text{si } F_{\beta_2=0} \geq c_\alpha. \end{cases}$$

Ici,  $F_{\beta_2=0}$  est la statistique de test en  $F$  habituelle pour l'hypothèse nulle  $\beta_2 = 0$ , et  $c_\alpha$  est la valeur critique pour un test de niveau  $\alpha$  qui est retournée par la distribution  $F(r, n-k)$ . Ainsi,  $\check{\beta}_1$  correspondra à l'estimateur contraint  $\tilde{\beta}_1$  lorsque le test en  $F$  ne permet pas de rejeter l'hypothèse selon laquelle les contraintes sont satisfaites, et correspondra à l'estimateur  $\hat{\beta}_1$  lorsque le test en  $F$  permet de rejeter cette hypothèse. Ceci est un exemple de ce que l'on nomme un **estimateur issu d'un test préliminaire** ou **estimateur i.t.p.**

Les estimateurs i.t.p. sont abondamment utilisés. Lorsque, par exemple, nous testons un quelconque caractère de spécification d'un modèle et décidons, sur la base des résultats fournis par le test, de quelle version nous allons faire l'estimation ou de la méthode d'estimation à employer, nous utilisons un estimateur i.t.p.. Malheureusement, les propriétés des estimateurs i.t.p. sont très difficiles à déceler dans la pratique. Ces difficultés apparaissent bien à partir de l'exemple que nous avons étudié. Supposons que les contraintes soient valides. L'estimateur qu'il faudrait utiliser serait alors l'estimateur contraint, c'est à dire  $\tilde{\beta}_1$ . Mais, dans  $\alpha\%$  des cas, le test en  $F$  rejettera à tort l'hypothèse nulle et  $\check{\beta}_1$  sera égal à l'estimateur non contraint  $\hat{\beta}_1$ . Ainsi  $\check{\beta}_1$  se révèle être moins efficace que  $\tilde{\beta}_1$  lorsque les contraintes sont valides. De plus, étant donné que la matrice de covariance estimée retournée par le progiciel de régression ne prend pas en compte le test préliminaire, les inférences pratiquées sur  $\check{\beta}_1$  pourraient être trompeuses.

D'autre part, lorsque les restrictions ne sont pas valides, nous pouvons utiliser l'estimateur non contraint  $\hat{\beta}_1$  ou pas. En fonction de la puissance du test en  $F$ ,  $\check{\beta}_1$  sera quelquefois égal à  $\tilde{\beta}_1$  et d'autres fois égal à  $\hat{\beta}_1$ . Il ne sera assurément pas non biaisé, parce que  $\tilde{\beta}_1$  est non biaisé, et il sera plus ou moins efficace (dans le sens de l'erreur quadratique moyenne) que l'estimateur non contraint. Les inférences sur  $\check{\beta}_1$  basées sur la matrice de covariance OLS estimée, que  $\check{\beta}_1$  soit égal à  $\tilde{\beta}_1$  ou à  $\hat{\beta}_1$ , pourraient être trompeuses, parce qu'elles ne tiennent pas compte du test préliminaire précédent.

Dans la pratique, il n'y a souvent pas grand chose à faire à propos des difficultés causées par les tests préliminaires sinon de reconnaître que ces tests sont source d'un facteur d'incertitude supplémentaire pour les problèmes de l'inférence statistique. Puisque  $\alpha$ , le niveau du test préliminaire, aura une influence sur les propriétés de  $\hat{\beta}_1$ , il est sans doute préférable d'essayer plusieurs valeurs de  $\alpha$ . Des niveaux de signification tels que .05 ne sont pas optimaux dans la plupart des cas, et il existe toute une littérature sur la façon de choisir des niveaux plus appropriés dans des cas bien déterminés; consulter, par exemple, Toyoda et Wallace (1976). Cependant, les véritables problèmes des tests préliminaires sont beaucoup plus délicats que celui dont nous avons discuté à titre d'exemple, ou que ceux étudiés dans la littérature. Chaque fois que l'on soumet un modèle à n'importe quel genre de test, le résultat obtenu influence peut-être la formulation définitive du modèle, et l'estimateur i.t.p. qui en découle devient donc encore plus complexe et il est très difficile d'imaginer comment on peut analyser cela de façon formelle.

Notre discours sur les estimateurs issus des tests préliminaires a été très concis. On trouvera des traitements plus explicites chez Fomby, Hill, et Johnson (1984, Chapitre 7), Judge, Hill, Griffiths, Lütkepohl, et Lee (1985, Chapitre 21), et Judge et Bock (1978). Dans le reste du livre nous ferons abstraction des problèmes engendrés par les tests préliminaires, non pas parce qu'ils sont marginaux, mais parce que dans la pratique, il sont généralement trop épineux.

### 3.8 CONCLUSION

Ce chapitre a présenté une introduction à certains thèmes importants: estimation de la matrice de covariance pour les estimations OLS; usage de telles matrices de covariance estimées dans l'élaboration des intervalles de confiance; idées de base des tests d'hypothèses; justification de l'utilisation des tests en  $t$  et en  $F$  pour mettre à l'épreuve des contraintes linéaires sur des modèles de régression linéaire; les trois principes classiques de tests d'hypothèses et leur application au modèle non linéaire. A plusieurs reprises, il nous a fallu rester imprécis et faire référence à des résultats concernant les propriétés asymptotiques des estimations par moindres carrés non linéaires que nous n'avons pas encore démontrées. Leur démonstration sera l'objet des deux chapitres suivants. Le Chapitre 4 s'intéresse aux idées de base de l'analyse asymptotique, entre autres la convergence, la normalité asymptotique, les Théorèmes de la Limite Centrale, les lois des grands nombres et l'usage des notations "grand- $O$ " et "petit- $o$ ". Puis le Chapitre 5 emploie ces concepts pour la démonstration de la convergence et de la normalité asymptotique des estimations par moindres carrés non linéaires des modèles de régression univariée et pour la dérivation des distributions asymptotiques des statistiques de test que nous avons rencontrées tout au long de ce chapitre. Nous y démontrerons

aussi un certain nombre de résultats asymptotiques dont nous apprécierons l'utilité par la suite.

## TERMES ET CONCEPTS

ellipse de confiance: relation avec les intervalles de confiance	pseudo- $F$ et pseudo- $t$
erreur de première espèce et de deuxième espèce	puissance d'un test
estimateur issu de test préliminaire (itp)	région critique, ou région de rejet
Fisher et distribution de $F$	région d'acceptation
fixes en échantillons répétés	régions de confiance et intervalles de confiance, exacts et approchés
hypothèse nulle et alternative	régresseurs artificiels
information a priori	régression de Gauss-Newton (GNR)
matrice de covariance	$R^2$ ajusté, ou $\bar{R}^2$
maximum de vraisemblance	statistique de test
$P$ marginal, ou niveau de signification marginale	Student et distribution de $t$
prétest	niveau d'un test
principe de Wald	test asymptotique
principe du multiplicateur de Lagrange (LM)	test convergent
principe du rapport de vraisemblance (LR)	test exact
	tests asymptotiquement équivalents
	valeur critique (pour une statistique de test)
	vecteur de score



# Chapitre 4

## Introduction aux Méthodes et à la Théorie Asymptotiques

### 4.1 INTRODUCTION

Une fois que l'on quitte le contexte des moindres carrés (linéaires) ordinaires avec des régresseurs fixes et des aléas normaux, il est souvent impossible, ou du moins irréalisable, d'obtenir des résultats statistiques exacts. Il est par conséquent nécessaire d'avoir recours à la **théorie asymptotique**, qui s'applique lorsque la taille de l'échantillon est infiniment grande. Les échantillons infinis ne sont pas disponibles dans cet univers fini, et seulement s'ils l'étaient il y aurait un contexte dans lequel la théorie asymptotique serait exacte. Naturellement, puisque les statistiques elles-mêmes seraient grandement inutiles si les échantillons étaient infiniment grands, la théorie asymptotique serait inutile si elle était exacte. Dans la pratique, la théorie asymptotique est utilisée selon les cas comme une approximation plus ou moins bonne.

La plupart du temps, croire que les résultats asymptotiques ont une certaine pertinence avec les données sur lesquelles on travaille relève davantage du vœu pieux que d'un résultat fondé. Malheureusement, des approximations plus précises ne sont disponibles que dans les cas les plus simples. A ce stade, il est probablement raisonnable d'indiquer que les principaux moyens de clarifier ces questions consistent à utiliser les expériences Monte Carlo, dont nous discuterons dans le dernier chapitre de ce livre. Puisqu'il n'est pas possible d'avoir recours à une expérience Monte Carlo à chaque fois que nous obtenons une statistique de test ou un ensemble d'estimations, une connaissance minutieuse de la théorie asymptotique est nécessaire dans l'état de l'art actuel et de la science économétrique. Le thème de ce chapitre consiste par conséquent à aborder l'étude de la théorie asymptotique qui sera utilisée dans le reste de ce livre. Toute cette théorie est fondamentalement basée sur les **lois des grands nombres** et sur les **théorèmes de la limite centrale**, et nous consacrerons par conséquent un temps considérable à discuter de ces résultats fondamentaux.

Dans ce chapitre, nous développons les idées de base et les concepts mathématiques nécessaires à la théorie asymptotique en économétrie. Nous commençons la prochaine section en traitant la notion fondamentale d'une suite infinie, d'éléments aléatoires ou non. La majeure partie de cette matière

devrait être connue de ceux qui ont étudié l'analyse réelle, mais il est utile d'y revenir parce que cela fournit directement des notions fondamentales de limites et de convergence, qui nous permettent d'établir et de démontrer une loi des grands nombres simples. Dans la Section 4.3, nous introduisons la notation “grand- $O$ ,” et “petit- $o$ ” et montrons comment l'idée d'une limite peut être utilisée pour obtenir des résultats plus précis et détaillés que ceux obtenus dans la Section 4.2. Les processus générateurs de données capables de générer des suites infinies de données sont introduits dans la Section 4.4, et ceci nécessite une brève discussion sur les processus stochastiques. La Section 4.5 est consacrée à la propriété de convergence d'un estimateur et montre comment cette propriété peut souvent s'exprimer à l'aide d'une loi des grands nombres. La normalité asymptotique est l'objet de la Section 4.6, et cette propriété est obtenue pour certains estimateurs simples grâce à un théorème de la limite centrale. Puis, dans la Section 4.7, nous fournissons, principalement pour l'intérêt d'une utilisation ultérieure, une série de définitions et de théorèmes, les derniers étant des lois des grands nombres et des théorèmes de la limite centrale beaucoup plus sophistiqués que ceux discutés dans le texte. De plus, nous présentons dans la Section 4.7 deux ensembles de conditions, l'un centré sur une loi des grands nombres, l'autre sur un théorème de la limite centrale, qui seront très utiles par la suite en tant que résumés des conditions de régularité nécessaires pour les résultats démontrés dans les chapitres suivants.

## 4.2 SUITES, LIMITES, ET CONVERGENCE

Le concept de l'infini suscite une perpétuelle fascination chez les mathématiciens. Un éminent mathématicien du vingtième siècle, Stanislaw Ulam, a écrit que l'évolution continue des notions variées de l'infini est une des premières forces motrices de la recherche en mathématiques (Ulam, 1976). Que cela soit vrai ou faux, les infinis apparemment irréalisables et certainement inaccessibles sont au cœur de la plupart des applications majeures et utiles des mathématiques utilisées de nos jours, et parmi elles l'économétrie.

La raison de l'utilisation généralisée de l'infini est qu'il peut fournir des approximations exploitables dans des circonstances où les résultats exacts sont difficiles voire impossibles à obtenir. L'opération mathématique cruciale qui conduit à ces approximations est le **passage à la limite**, la limite étant l'état où la notion d'infini apparaît. Les limites intéressantes peuvent être nulles, finies, ou infinies. Les limites nulles ou finies fournissent habituellement des approximations recherchées: des éléments difficiles à évaluer dans un contexte réaliste et fini sont remplacés par leurs limites comme approximation.

La **suite** est la construction mathématique possédant une limite de loin la plus fréquente. Une suite est une collection infinie d'éléments, tels des nombres, des vecteurs, des matrices, ou plus généralement des objets mathématiques, de sorte que, par définition, elle ne peut pas se représenter

dans le monde physique réel. Certaines suites sont néanmoins très familières. Considérons la plus célèbre de toutes: la suite

$$\{1, 2, 3, \dots\}$$

des entiers naturels. Ceci est peut-être un exemple simpliste, mais qui exhibe la plupart des propriétés importantes que peuvent posséder les suites. La première d'entre elles est qu'une suite doit comprendre une **règle** qui la définit. Dans le monde physique, nous pouvons définir une suite en indiquant ou en désignant tous ses membres, à condition qu'ils soient en nombre fini, mais cela est impossible pour une suite infinie. Par conséquent, il doit exister une méthode pour générer les éléments d'une suite, et la règle remplit cette fonction. Pour les entiers naturels, la règle est simple: on part de n'importe quel élément de la suite vers son **terme suivant** en ajoutant 1.

Cette dernière remarque illustre une autre propriété des suites comparativement à d'autres collections infinies: une suite est **ordonnée**. Ainsi, nous pouvons parler du premier élément de la suite, du deuxième, du troisième, et des suivants, et la règle qui définit la suite doit être capable de générer le  $n^{\text{ième}}$  élément, pour n'importe quel entier positif  $n$ .

La suite des entiers naturels est dans un certain sens le modèle de toutes les suites, parce qu'elle exprime la notion de succession des éléments, à savoir la notion de passage d'un élément à un autre, son successeur. Formellement, une suite peut être définie comme une application de l'ensemble ou de la suite des entiers naturels dans un autre ensemble quelconque, par exemple celui des nombres réels ou l'ensemble des matrices de dimension  $n \times m$ , à partir de laquelle les éléments de la suite sont construits. Cette application comprend la règle qui définit la suite, puisqu'elle associe à toute entier  $n$  le  $n^{\text{ième}}$  élément de la suite. Si l'action de cette application peut s'exprimer simplement, elle fournit alors une notation très pratique pour les suites. Nous notons simplement entre accolades le  $n^{\text{ième}}$  élément de la suite, et, juste comme nous le ferions avec le symbole de sommation, et pouvons indiquer l'étendue de la suite. Ainsi, la suite des entiers naturels peut se noter  $\{n\}$  ou  $\{n\}_{n=1}^{\infty}$ . Notons que le "premier" élément de la suite n'est pas forcément indicé par 1: nous pouvons parfaitement considérer la suite  $\{n\}_{n=m}^{\infty}$  des entiers supérieurs ou égaux à  $m$ .

Comme nous l'avons indiqué précédemment, nous nous focaliserons tout d'abord sur des suites ayant des limites finies pour l'analyse asymptotique. La suite des entiers naturels ne possède pas une telle caractéristique, mais il n'est pas difficile de trouver des suites pour lesquelles c'est le cas. Par exemple,  $\{1/n\}$ ,  $\{e^{-n}\}$ ,  $\{1/\log n\}$ , et  $\{n^{-2}\}$  ont toutes des limites nulles. Les suites suivantes, par contre, ont toutes des limites finies non nulles:

$$\left\{ \frac{n}{n+1} \right\}, \quad \left\{ n \sin\left(\frac{1}{n}\right) \right\}, \quad \left\{ 1 + \frac{1}{n} \right\}, \quad \left\{ n(y^{1/n} - 1) \right\}.$$

Si le lecteur ne peut pas calculer les valeurs de ces limites, il serait très utile d'apprendre à le faire. La dernière n'est pas très facile, mais, comme nous le verrons dans le Chapitre 14, elle est parfois très utile dans la modélisation économétrique.

Les limites de ces suites sont les limites **lorsque  $n$  tend vers l'infini**. Nous pouvons parfois dire à la place que  $n$  devient infiniment grand, ou, par abus de langage, simplement grand. Une autre possibilité consiste à discuter de la limite pour un  $n$  grand. Dans tous les cas, la signification devrait être claire. La définition formelle de la limite d'une **suite réelle**, à savoir une suite dont les éléments sont des nombres réels, est comme suit:

*Définition 4.1.*

La suite réelle  $\{a_n\}$  a pour limite le nombre réel  $a$ , ou converge vers  $a$ , si pour n'importe quel  $\varepsilon$  positif, aussi petit soit-il, il est possible de trouver un entier positif  $N$  tel que pour tous les entiers  $n$  supérieurs à  $N$ ,  $|a_n - a| < \varepsilon$ .

Autrement dit, lorsque  $n$  devient grand, nous pouvons toujours sélectionner un point de la suite au-delà duquel la différence entre les éléments de la suite et la limite est inférieure à n'importe quel seuil de tolérance prédéfini.

Dans la Définition 4.1, nous avons employé le terme important *converge*. Si les suites ont des limites, nous disons qu'elles **convergent** vers ces limites. Une suite qui converge est dite **convergente**. De façon alternative, une suite qui n'a pas de limite **diverge**, ou est **divergente**, si les valeurs absolues des éléments de la suite augmentent sans borne quand  $n$  devient plus grand. Il existe d'autres possibilités, en particulier si les éléments de la suite sont des objets plus compliqués que des nombres réels, telles les matrices. La convergence d'une suite ne peut être abordée que si les éléments de la suite appartiennent à un ensemble sur lequel peut se définir l'idée de **distance**, du fait de la nécessité de savoir si les éléments se rapprochent de plus en plus de la limite quand  $n$  tend vers l'infini. Ainsi, pour qu'une suite de vecteurs ou de matrices converge, nous devons être capables de dire si deux vecteurs ou matrices sont proches ou non, compte tenu d'un niveau de tolérance donné. Pour les vecteurs, ceci est facile: nous pouvons utiliser le nombre réel non négatif  $\|\mathbf{v}_1 - \mathbf{v}_2\|$ , la distance Euclidienne, comme mesure de la distance entre les deux vecteurs  $\mathbf{v}_1$  et  $\mathbf{v}_2$  dans un espace Euclidien. Pour des matrices, une mesure comparable est la norme de la différence entre deux matrices; consulter l'Annexe A pour la définition de la norme d'une matrice, et se souvenir également que la notation utilisée est habituellement  $\|\cdot\|$ , tout comme pour la distance Euclidienne. Cet élément aidera la concision dans l'écriture des définitions.

La discussion générale de la distance est l'objet de la discipline mathématique appelée topologie. La convergence et les limites ne sont définies que sur des **espaces topologiques**. L'ensemble des nombres réels et des espaces Euclidiens possède ce que l'on appelle des **topologies naturelles**. Il

s'agit de celles utilisées dans la Définition 4.1 pour des nombres réels et qui peuvent l'être dans l'extension de la définition à des vecteurs ou des matrices basée sur l'usage de la norme Euclidienne  $\|\cdot\|$ . Ces topologies sont en fait tellement naturelles que, lorsque nous discutons de la convergence de suites réelles, vectorielles ou matricielles, il est souvent inutile d'explicitier les concepts topologiques. Ceci n'est malheureusement pas le cas quand nous considérons des suites de **variables aléatoires**.<sup>1</sup> Pour compliquer davantage les choses, il n'existe pas de topologie naturelle unique pour les variables aléatoires; au moins trois ou quatre sont régulièrement utilisées.

Il n'est pas nécessaire et serait inopportun dans un ouvrage de ce genre de donner des définitions formelles des différentes topologies utilisées avec les variables aléatoires, aussi ne le ferons-nous pas. Pour les lecteurs qui s'en trouvent insatisfaits, nous pouvons recommander les deux livres de Billingsley (1968, 1979). Tout ce qu'il nous faut savoir sur une topologie est si une suite donnée converge ou non pour cette topologie, de sorte que les définitions des différentes sortes de convergence pour les variables aléatoires que nous donnerons suffiront.

Vraisemblablement la sorte de convergence stochastique la plus utile, c'est-à-dire la convergence pour des variables aléatoires, est la **convergence en probabilité**. Nous commençons par la définition formelle:

*Définition 4.2.*

La suite  $\{\mathbf{a}_n\}$  de variables aléatoires réelles ou vectorielles tend en probabilité vers la variable aléatoire limite  $\mathbf{a}$  si pour tout  $\varepsilon$  et  $\delta > 0$  il existe un  $N$  tel que pour tout  $n > N$ ,

$$\Pr(\|\mathbf{a}_n - \mathbf{a}\| > \varepsilon) < \delta. \quad (4.01)$$

Dans ce cas,  $\mathbf{a}$  est appelée la **limite en probabilité** ou simplement la **plim** de la suite  $\{\mathbf{a}_n\}$ . On écrit

$$\text{plim}_{n \rightarrow \infty} \mathbf{a}_n = \mathbf{a} \quad \text{ou} \quad \mathbf{a}_n \xrightarrow{p} \mathbf{a}.$$

Notons l'absence d'accolades dans ces dernières expressions. Notons également que, pour des variables aléatoires réelles, la norme Euclidienne  $\|\cdot\|$  se simplifie en la valeur absolue ordinaire  $|\cdot|$ .

La condition (4.01) indique, de fait, que pour n'importe quel niveau de tolérance donné  $\varepsilon$ , nous pouvons atteindre un élément de la suite tel que, au-delà de cet élément, la probabilité de trouver un écart entre un élément de la suite et la variable aléatoire limite supérieur au niveau de tolérance est inférieure à un autre niveau de tolérance prédéfini  $\delta$ . Notons que, bien que la limite en probabilité de la définition précédente  $\mathbf{a}$  soit une variable aléatoire

<sup>1</sup> Pour une discussion du sens du terme **variable aléatoire**, consulter l'Annexe B.

(ou un vecteur de variables aléatoires) il peut s'agir en fait d'un nombre ou d'un vecteur ordinaire, auquel cas elle est dite **non stochastique**, ou constante.

Un exemple rebattu d'une limite en probabilité non stochastique est celui de la limite de la suite des proportions des faces dans une série de lancements indépendants d'une pièce de monnaie sans biais. Il est utile de démontrer formellement que la limite en probabilité est en effet un demi, puisque ceci nous donnera l'opportunité de présenter certaines techniques de démonstration utiles et d'acquérir une certaine intuition sur la manière dont les limites en probabilité diffèrent des limites ordinaires.

Alors, pour chaque lancement de pièce, définissons une variable aléatoire  $y_t$  égale à 1 si le résultat est face et 0 si le résultat est pile. Ceci signifie que  $\{y_t\}$  est une suite de variables aléatoires, à condition d'imaginer que l'expérience se répète à l'infini. Alors, après  $n$  lancers, la proportion de faces est juste

$$a_n \equiv \frac{1}{n} \sum_{t=1}^n y_t, \quad (4.02)$$

et, naturellement,  $a_n$  est un nombre réel (de fait rationnel) appartenant à l'intervalle  $[0,1]$ . L'expression (4.02) définit une autre suite de variables aléatoires  $\{a_n\}$ , dont la limite en probabilité est précisément ce que nous désirons calculer.

Nous calculons tout d'abord l'espérance et la variance de  $a_n$ . Le fait que la pièce soit sans biais signifie que, pour tout  $t$ ,  $\Pr(y_t = 1) = \frac{1}{2}$  et  $\Pr(y_t = 0) = \frac{1}{2}$ . Du fait de la linéarité du calcul des espérances et de l'identité des espérances,

$$E(a_n) = \frac{1}{n} \sum_{t=1}^n E(y_t) = E(y_t) = \frac{1}{2}.$$

Le calcul de la variance est légèrement plus délicat. Nous voyons que

$$\text{Var}(a_n) = E(a_n - E(a_n))^2 = E(a_n - \frac{1}{2})^2 = E\left(\frac{1}{n} \sum_{t=1}^n \left(y_t - \frac{1}{2}\right)\right)^2.$$

Avant de procéder, définissons une nouvelle suite de variables aléatoires indépendantes  $\{z_t\}$  d'après la règle  $z_t = y_t - \frac{1}{2}$ . Comme  $E(z_t) = 0$ ,  $\{z_t\}$  est une **suite centrée**, par laquelle nous signifions que chaque élément de la suite a une espérance nulle, nous trouvons alors que

$$a_n - E(a_n) = \frac{1}{n} \sum_{t=1}^n y_t - \frac{1}{2} = \frac{1}{n} \sum_{t=1}^n \left(y_t - \frac{1}{2}\right) = \frac{1}{n} \sum_{t=1}^n z_t.$$

Nous pouvons définir une autre suite centrée  $\{b_n\}$  d'après la règle

$$b_n = a_n - E(a_n) = a_n - \frac{1}{2} \quad (4.03)$$

et nous voyons que  $\text{Var}(a_n) = \text{Var}(b_n)$ . Nous voyons également que

$$b_n = \frac{1}{n} \sum_{t=1}^n z_t. \quad (4.04)$$

Comme les  $z_t$  sont mutuellement indépendantes,

$$\text{Var}(b_n) = \frac{1}{n^2} \sum_{t=1}^n \text{Var}(z_t) = n^{-1} \text{Var}(z_t).$$

Il est simple de voir que

$$\text{Var}(z_t) = \left(\frac{1}{2}\right)^2 (\text{Pr}(z_t = \frac{1}{2})) + \left(-\frac{1}{2}\right)^2 (\text{Pr}(z_t = -\frac{1}{2})) = \frac{1}{4},$$

d'où découle que

$$\text{Var}(a_n) = \text{Var}(b_n) = \frac{1}{4n}. \quad (4.05)$$

Ce résultat est crucial pour notre objectif, puisqu'il implique que

$$\lim_{n \rightarrow \infty} \text{Var}(a_n) = 0, \quad (4.06)$$

et il est ainsi intuitivement évident que la limite de la suite  $\{a_n\}$  est non aléatoire. Mais de façon formelle, il reste quelques étapes à franchir pour établir le résultat nécessaire que, en appliquant (4.01) aux circonstances actuelles, pour tous  $\varepsilon$  et  $\delta > 0$ , il existe un  $N$  tel que

$$\text{Pr}(|a_n - \frac{1}{2}| > \varepsilon) < \delta \quad \text{pour tout } n > N. \quad (4.07)$$

Le passage de (4.06) à (4.07) s'opère principalement par l'usage de l'**inégalité de Chebyshev** (consulter l'Annexe B). Cette inégalité nous indique que si une variable aléatoire  $y$  d'espérance nulle a une variance  $V$ , alors pour tout nombre positif  $\alpha$

$$\text{Pr}(|y| > \alpha) < \frac{V}{\alpha^2}.$$

Si nous appliquons cette formule à la variable  $b_n$ , alors à partir de (4.05)

$$\text{Pr}(|b_n| > \varepsilon) < \frac{1}{4n\varepsilon^2}.$$

A partir de la définition (4.03) de  $b_n$ , ceci signifie que

$$\text{Pr}(|a_n - \frac{1}{2}| > \varepsilon) < \frac{1}{4n\varepsilon^2}.$$

Ainsi (4.07) sera vraie si nous choisissons, pour un  $\varepsilon$  donné, une valeur critique de  $N$  égale à l'entier supérieur le plus proche de  $(4\varepsilon^2\delta)^{-1}$ . La convergence en probabilité de  $\{a_n\}$  vers  $\frac{1}{2}$ ,

$$\text{plim}_{n \rightarrow \infty} a_n = \frac{1}{2}, \quad (4.08)$$

est, finalement, rigoureusement démontrée.

Comme produit dérivé de la démonstration précédente, nous voyons que l'inégalité de Chebyshev indique que toute suite centrée dont la variance tend vers zéro tend vers zéro en probabilité. Supposons que la suite  $\{y_n\}$  soit centrée, que  $v_n = \text{Var}(y_n)$ , et que  $v_n \rightarrow 0$  quand  $n \rightarrow \infty$ . D'après la Définition 4.1 de la limite d'une suite, cette dernière hypothèse signifie que pour tout  $\eta > 0$ , nous pouvons trouver un  $N(\eta)$  tel que  $v_n < \eta$  pour tout  $n > N(\eta)$ . Nous examinons ensuite la probabilité qui apparaît dans la Définition 4.2 de la convergence en probabilité. Pour un  $\varepsilon > 0$  quelconque,

$$\Pr(|y_n| > \varepsilon) < v_n \varepsilon^{-2}, \quad \text{d'après l'inégalité de Chebyshev.}$$

Considérons maintenant la grandeur critique  $N(\delta\varepsilon^2)$  pour tout  $\delta$  positif, telle que, pour tout  $n > N(\delta\varepsilon^2)$ ,  $v_n < \delta\varepsilon^2$ . Pour un tel  $n$  nous trouvons que

$$\Pr(|y_n| > \varepsilon) < \delta\varepsilon^2 \varepsilon^{-2} = \delta,$$

exactement comme cela était demandé dans la Définition 4.2.

Le résultat (4.08) est un premier exemple de ce qui est appelé une **loi des grands nombres**. L'idée sous-jacente de cet exemple est en fait la même que celle sous-jacente à toutes les lois des grands nombres, et il est par conséquent utile de s'attarder pour la considérer de façon plus approfondie. Les éléments de la suite  $\{a_n\}$  sont tous des *moyennes*, la proportion du nombre de fois qu'une pièce est tombée "face" au cours des lancers. Quand nous augmentons la valeur de  $n$ , nous nous attendons à ce que la moyenne calculée à partir d'un échantillon aléatoire de taille  $n$  soit une mesure de plus en plus précise d'une certaine grandeur. Ici la grandeur est juste le nombre  $\frac{1}{2}$ , l'inverse du nombre de côtés que comporte une pièce. Si nous lançons des dés à la place de pièces de monnaie, nous nous attendrions à obtenir  $\frac{1}{6}$  au lieu de  $\frac{1}{2}$ .

Evidemment, la même sorte de résultat prévaudra si nous mesurons une grandeur plus intéressante pour les économistes qu'une proportion de piles ou de faces. Nous pourrions, par exemple, être intéressés par la mesure de la proportion des individus qui possèdent leur propre habitation, disons  $\alpha$ , pour un certain groupe de gens. En supposant que nous pouvons trouver une méthode d'échantillonnage aléatoire à partir de la population pertinente (ce qui est souvent loin d'être une tâche facile), nous pourrions demander à toutes les personnes de l'échantillon si elles sont propriétaires ou pas de leur habitation. Chaque réponse peut alors être traitée exactement de la même façon que le lancement d'une pièce; la variable aléatoire  $y_t$  peut prendre la valeur 1 pour



ceux qui répondent être propriétaires de leur habitation et 0 pour les autres. La loi des grands nombres nous indique alors que, quand le nombre de réponses devient important, nous devrions nous attendre à ce que la proportion des individus qui sont propriétaires de leur habitation,  $a_n \equiv n^{-1} \sum_{t=1}^n y_t$ , converge vers la véritable valeur  $\alpha$ . Intuitivement, la précision accrue provient du fait que chaque lancement successif apporte une information supplémentaire sur  $\alpha$ .

A ce stade, nous pouvons introduire quelques termes standards. La forme de la loi des grands nombres que nous avons démontrée pour l'exemple du lancé d'une pièce, dans lequel nous montrions que la limite en probabilité de la proportion de faces tend vers un demi, est appelée **loi faible des grands nombres**, parce que la sorte de convergence démontrée est la convergence en probabilité. Il existe des **lois fortes des grands nombres**, qui utilisent, comme le terme le suggère, une notion plus forte de la convergence des variables aléatoires, appelée **convergence presque sûre**. En voici la définition:

*Définition 4.3.*

La suite  $\{\mathbf{a}_n\}$  de variables aléatoires réelles ou vectorielles  $\mathbf{a}_n$  est dite converger presque sûrement (a.s.) vers une variable aléatoire limite  $\mathbf{a}$  si

$$\Pr\left(\lim_{n \rightarrow \infty} \mathbf{a}_n = \mathbf{a}\right) = 1. \quad (4.09)$$

Nous notons

$$\lim_{n \rightarrow \infty} \mathbf{a}_n = \mathbf{a} \text{ a.s.} \quad \text{ou} \quad \mathbf{a}_n \xrightarrow{\text{a.s.}} \mathbf{a} \quad \text{ou} \quad \mathbf{a}_n \rightarrow \mathbf{a} \text{ a.s.},$$

et  $\mathbf{a}$  est appelée la **limite presque sûre** de  $\{\mathbf{a}_n\}$ .

Une compréhension totale de la définition précédente nécessite une certaine connaissance plus approfondie de la théorie probabiliste que ce que nous attendons de nos lecteurs, aussi n'en discuterons-nous plus par la suite. De façon similaire, une démonstration de la loi forte des grands nombres, même dans le cas d'un simple lancé de pièce, dépasse la portée de ce livre. Une démonstration rigoureuse et originale est disponible dans le premier chapitre de Billingsley (1979), et le traitement classique est donné dans Feller (1968). Dans la suite de ce livre, nous nous contenterons d'employer les lois faibles des grands nombres, même si les lois fortes sont disponibles, puisque la distinction entre les deux formes n'a pas d'implication pratique pour l'économétrie.

Une troisième forme de convergence stochastique est appelée la **convergence en distribution**, ou parfois **convergence en loi**, terme qui s'inspire du fait que la distribution d'une variable aléatoire est appelée sa **loi**. Cette convergence en distribution est habituellement démontrée dans un théorème de la limite centrale, comme nous le verrons plus tard dans ce chapitre.

*Définition 4.4.*

La suite  $\{\mathbf{a}_n\}$  de variables aléatoires réelles vectorielles  $\mathbf{a}_n$  converge en distribution vers une variable aléatoire limite  $\mathbf{a}$  si

$$\lim_{n \rightarrow \infty} \Pr(\mathbf{a}_n \leq \mathbf{b}) = \Pr(\mathbf{a} \leq \mathbf{b}) \quad (4.10)$$

pour tout nombre réel ou vecteur  $\mathbf{b}$  tels que la fonction de distribution limite  $\Pr(\mathbf{a} \leq \mathbf{b}')$  est continue par rapport à  $\mathbf{b}'$  en  $\mathbf{b}' = \mathbf{b}$ . On écrit :

$$\mathbf{a}_n \xrightarrow{D} \mathbf{a}.$$

Les inégalités dans (4.10) doivent être interprétées, dans le cas de variables aléatoires vectorielles, comme s'appliquant à chaque élément des vecteurs séparément. Ceci est exactement comme dans la définition formelle de l'Annexe B de la distribution de probabilité jointe d'un ensemble de variables aléatoires. La contrainte que  $\Pr(\mathbf{a} \leq \mathbf{b}')$  soit continue par rapport à  $\mathbf{b}'$  en  $\mathbf{b}' = \mathbf{b}$  dans la définition est évidemment inutile dans le cas des distributions continues, pour lesquelles la définition nécessite simplement que les fonctions de distribution des  $\mathbf{a}_n$  convergent en tout point vers la fonction de distribution de  $\mathbf{a}$ . Mais si la variable aléatoire limite n'est pas stochastique, cette contrainte est nécessaire. La raison est que la fonction de distribution (parfois appelée c.d.f.) est nécessairement discontinue dans ce cas. Considérons l'exemple suivant.

Soit  $x_n$  une variable aléatoire obéissant à une  $N(0, n^{-1})$ . Clairement,  $\{x_n\}$  converge vers zéro quelle que soit la définition de la convergence, et, en particulier, pour la convergence en distribution. Du fait que la variance de  $x_n$  est  $n^{-1}$ , sa c.d.f. est  $\Phi(n^{1/2}x)$ , dans le sens où

$$\Pr(x_n < x) = \Phi(n^{1/2}x) \quad \text{pour tout réel } x.$$

Ici  $\Phi(\cdot)$  est la c.d.f. de la normale centrée réduite, notée  $N(0, 1)$ ; consulter l'Annexe B pour les détails. Pour un  $x$  fixé, nous avons

$$\lim_{n \rightarrow \infty} n^{1/2}x = \begin{cases} \infty & \text{si } x > 0; \\ 0 & \text{si } x = 0; \\ -\infty & \text{si } x < 0. \end{cases}$$

Puisque

$$\lim_{x \rightarrow \infty} \Phi(x) = 1, \quad \lim_{x \rightarrow -\infty} \Phi(x) = 0, \quad \text{et} \quad \Phi(0) = \frac{1}{2},$$

nous obtenons

$$\lim_{n \rightarrow \infty} \Pr(x_n < x) = \begin{cases} 0 & \text{si } x < 0; \\ \frac{1}{2} & \text{si } x = 0; \\ 1 & \text{si } x > 0. \end{cases}$$

La limite précédente coïncide quasiment avec la c.d.f. d'une variable "aléatoire"  $x_0$  qui est de fait toujours égale à zéro. Cette c.d.f., qui correspond à ce que l'on appelle une **distribution dégénérée** concentrée en 0, est

$$\Pr(x_0 < x) = \begin{cases} 0 & \text{si } x \leq 0; \\ 1 & \text{si } x > 0. \end{cases}$$

C'est seulement en  $x = 0$  que la limite des c.d.f. des  $x_n$  n'est pas égale à la c.d.f. de la variable aléatoire constante. Mais c'est précisément dans ce cas que cette dernière est nécessairement discontinue, d'où l'exception explicitement formulée dans la définition. Une c.d.f. qui comporte des discontinuités en certains points est dite comporter des **atomes** en ces points. Notons qu'une c.d.f. avec des atomes peut parfaitement être la limite d'une suite de c.d.f. qui n'ont aucun atome et qui par conséquent sont partout continues.

Nous concluons cette section en établissant sans démonstration les relations entre les trois sortes de convergence stochastique introduites jusqu'ici. La convergence presque sûre est, comme le terme *forte* associé à la loi des grands nombres le suggère, la sorte la plus forte. Si  $\{a_n\}$  converge presque sûrement vers une variable limite  $a$ , alors elle converge aussi vers  $a$  en probabilité et en distribution. La convergence en probabilité, bien qu'elle n'implique pas nécessairement la convergence presque sûre, implique la convergence en distribution. La convergence en distribution est la sorte de convergence la plus faible des trois et n'implique pas nécessairement l'une des deux autres.

### 4.3 TAUX DE CONVERGENCE

Dans la dernière section nous avons couvert un champ important, tellement important que nous avons abordé, même succinctement, tous les thèmes purement mathématiques importants dont nous discuterons dans ce chapitre. Il reste à enrichir la discussion de certaines matières et à appliquer notre théorie à la statistique et à l'économétrie. Le sujet de cette section concerne les **taux de convergence**. A travers ce sujet, nous introduirons des notations très importantes, appelées **notations  $O, o$** , qui se lisent "notation grand- $O$ ", et "notation petit- $o$ ." Ici  $O$  et  $o$  indiquent l'ordre et sont souvent considérés comme des **symboles d'ordre**. De manière approximative, nous dirons qu'une certaine quantité est, disons,  $O(x)$ , lorsqu'elle est du même ordre, asymptotiquement, que la quantité  $x$ , alors qu'elle sera  $o(x)$  quand elle sera d'un ordre inférieur à la quantité  $x$ . La signification précise de tout cela sera indiqué par la suite.

Dans la section précédente, nous discutons de la variable aléatoire  $b_n$  en détail, et comprenions à partir de (4.05) que sa variance convergeait vers zéro, parce qu'elle était proportionnelle à  $n^{-1}$ . Ceci implique que la suite converge en probabilité vers zéro, et que les moments d'ordre supérieur de  $b_n$ , le troisième, le quatrième, et ainsi de suite, tendent également vers zéro lorsque  $n \rightarrow \infty$ . Un calcul quelque peu astucieux, auquel les lecteurs intéressés sont invités à se livrer, révèle que le quatrième moment de  $b_n$  est

$$E(b_n^4) = \frac{3}{16}n^{-2} - \frac{1}{8}n^{-3}, \quad (4.11)$$

à savoir la somme de deux termes, un proportionnel à  $n^{-2}$  et l'autre à  $n^{-3}$ . Le troisième moment de  $b_n$ , comme le premier, est nul, simplement du fait que la variable aléatoire est **symétrique** autour de zéro ce qui implique que

tous les moments d'ordre impair s'annulent. Ainsi, le second moment de  $b_n$ , le troisième et le quatrième convergent vers zéro, mais à des *taux* différents. A nouveau, les deux termes dans le quatrième moment (4.11) convergent à des taux différents, et c'est le terme proportionnel à  $n^{-2}$  qui est le plus important asymptotiquement.

Le mot "asymptotiquement" a été utilisé ici dans un sens légèrement plus large que précédemment. Dans la Section 4.1, nous disions que la théorie asymptotique traite des limites par rapport à un certain indice, habituellement la taille d'échantillon en économétrie, qui tend vers l'infini. Ici nous sommes intéressés par les taux de convergence plutôt que par les limites en tant que telles. Les limites peuvent être utilisées pour déterminer les taux de convergence des suites aussi bien que leurs valeurs limites: ces taux de convergence peuvent être définis comme les limites d'autres suites. Par exemple, dans la comparaison de  $n^{-2}$  et  $n^{-3}$ , l'autre suite qui nous intéresse est la suite du *ratio* de  $n^{-3}$  sur  $n^{-2}$ , à savoir la suite  $\{n^{-1}\}$ . Cette dernière suite a une limite nulle, et par conséquent, asymptotiquement, nous pouvons considérer  $n^{-3}$ , ou n'importe quelle quantité proportionnelle, comme nulle par rapport à  $n^{-2}$ , ou n'importe quelle quantité proportionnelle. Tout ceci peut s'exprimer par la notation petit- $o$ , qui exprime ce que l'on appelle **relation de petit-ordre**: nous notons  $n^{-3} = o(n^{-2})$ , ce qui signifie que  $n^{-3}$  est d'un ordre plus faible que  $n^{-2}$ . En général, nous avons la définition suivante:

*Définition 4.5.*

Si  $f(\cdot)$  et  $g(\cdot)$  sont deux fonctions réelles de la variable positive entière  $n$ , alors la notation

$$f(n) = o(g(n)) \quad [\text{de façon facultative, quand } n \rightarrow \infty]$$

signifie que

$$\lim_{n \rightarrow \infty} \left( \frac{f(n)}{g(n)} \right) = 0.$$

Nous pouvons dire que  $f(n)$  est d'un ordre plus faible que  $g(n)$  asymptotiquement ou quand  $n$  tend vers l'infini.

Notons que  $g(n)$  elle-même peut avoir toute sorte de comportement quand  $n \rightarrow \infty$ . Elle peut posséder une limite ou pas, et si c'est le cas, cette limite peut être nulle, finie et non nulle, ou infinie. Ce qui est important est la *comparaison* opérée par le ratio. Le plus souvent  $g(n)$  est une puissance de  $n$ , positive, négative, ou nulle. Dans le dernier cas, puisque  $n^0 = 1$  pour tout  $n$ , nous devrions écrire  $f(n) = o(1)$ , et ceci signifierait d'après la définition que

$$\lim_{n \rightarrow \infty} \left( \frac{f(n)}{1} \right) = \lim_{n \rightarrow \infty} (f(n)) = 0;$$

autrement dit, que  $f(n)$  tend vers zéro quand  $n$  tend vers l'infini. Mais si nous disons que  $f(n) = o(n^{-1})$ , par exemple, ou que  $f(n)$  est  $o(n^{-1})$ , nous

signifions que  $f(n)$  tend vers zéro *plus vite* que  $n^{-1}$ . Nous pourrions également dire que  $f(n)$  est  $o(n)$ , et nous ne saurions pas alors si  $f(n)$  possède une limite quand  $n \rightarrow \infty$ . Mais nous savons que si  $f(n)$  tend vers l'infini, c'est moins rapidement que  $n$ .

La notation grand- $O$ , qui exprime la **relation d'ordre identique**, est plus précise que la notation petit- $o$ , puisqu'elle nous indique le taux le plus fort auquel les quantités peuvent varier avec  $n$ . En voici la définition.

*Définition 4.6.*

Si  $f(\cdot)$  et  $g(\cdot)$  sont deux fonctions réelles de la variable positive entière  $n$ , alors la notation

$$f(n) = O(g(n))$$

signifie qu'il existe une constante  $K > 0$ , indépendante de  $n$ , et un entier positif  $N$  tels que  $|f(n)/g(n)| < K$  pour tout  $n > N$ .

Nous disons que  $f(n)$  et  $g(n)$  sont du même ordre asymptotiquement ou quand  $n \rightarrow \infty$ . Une fois encore c'est le *ratio* de  $f(n)$  sur  $g(n)$  qui est en cause. La définition *n'exclut pas* la possibilité que la limite du ratio soit nulle, de sorte que l'expression verbale "du même ordre" peut être trompeuse.

Une autre relation évite cette incertitude: nous l'appelons l'**égalité asymptotique**.

*Définition 4.7.*

Si  $f(\cdot)$  et  $g(\cdot)$  sont des fonctions réelles de la variable positive entière  $n$ , alors elles sont asymptotiquement égales si

$$\lim_{n \rightarrow \infty} \left( \frac{f(n)}{g(n)} \right) = 1.$$

Nous notons  $f(n) \stackrel{a}{=} g(n)$ . La notation standard pour cette relation (en dehors de l'économétrie) n'est pas  $\stackrel{a}{=}$ , mais  $\sim$ . Puisque le symbole  $\sim$  est utilisé pour désigner la distribution d'une variable aléatoire, cela justifiera l'usage de l'autre symbole dans ce livre.

L'égalité asymptotique évite la difficulté à laquelle nous avons fait allusion en connexion avec le grand- $O$  ou relation du même ordre, aux dépens d'une condition plus forte. Contrairement à l'égalité asymptotique, la relation grand- $O$  ne nécessite pas que le ratio  $f(n)/g(n)$  ait une quelconque limite. Il peut en avoir une, mais il peut également varier perpétuellement entre des bornes.

Les relations définies plus tôt sont consacrées aux suites réelles non stochastiques. Les relations dites **relations d'ordre stochastiques** sont d'un intérêt encore plus grand pour les économètres. Elles sont parfaitement analogues aux relations définies plus haut, mais reposent par contre sur l'une ou l'autre forme de la convergence stochastique. Formellement:

*Définition 4.8.*

Si  $\{a_n\}$  est une suite de variables aléatoires, et  $g(n)$  une fonction réelle d'un argument positif entier  $n$ , alors la notation  $a_n = o_p(g(n))$  signifie que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{a_n}{g(n)} \right) = 0.$$

De façon similaire, la notation  $a_n = O_p(g(n))$  signifie qu'il existe une constante  $K$  telle que, pour tout  $\varepsilon > 0$ , il existe un entier positif  $N$  tel que

$$\Pr \left( \left| \frac{a_n}{g(n)} \right| > K \right) < \varepsilon \quad \text{pour tout } n > N.$$

Si  $\{b_n\}$  est une autre suite de variables aléatoires, la notation  $a_n \stackrel{a}{=} b_n$  signifie que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{a_n}{b_n} \right) = 1.$$

Des définitions comparables peuvent étre établies pour la convergence presque sûre et la convergence en distribution, mais nous ne les utiliserons pas. En fait, après cette section nous ne nous embarrasserons plus de l'indice  $p$  dans les symboles d'ordre stochastiques, parce qu'il sera toujours évident de savoir si les variables en question sont aléatoires. Quand elles le sont,  $O(\cdot)$  et  $o(\cdot)$  devraient être interprétés comme  $O_p(\cdot)$  et  $o_p(\cdot)$ .

Les symboles d'ordre sont très simples à manipuler, et nous présentons maintenant quelques règles de manipulations. Pour faire simple, nous nous restreignons aux fonctions  $g(n)$  qui sont simplement des puissances de  $n$ , les seules que nous utilisons dans ce livre. Les règles pour l'addition et la soustraction sont

$$\begin{aligned} O(n^p) \pm O(n^q) &= O(n^{\max(p,q)}); \\ o(n^p) \pm o(n^q) &= o(n^{\max(p,q)}); \\ O(n^p) \pm o(n^q) &= O(n^p) \quad \text{si } p \geq q; \\ O(n^p) \pm o(n^q) &= o(n^q) \quad \text{si } p < q. \end{aligned}$$

Les règles pour la multiplication, et aussi pour la division, sont

$$\begin{aligned} O(n^p) O(n^q) &= O(n^{p+q}); \\ o(n^p) o(n^q) &= o(n^{p+q}); \\ O(n^p) o(n^q) &= o(n^{p+q}). \end{aligned}$$

Bien que ces règles gèrent de manière satisfaisante tous les cas simples, elles ne gèrent pas les cas où des quantités toutes d'un certain ordre sont additionnées. De tels cas surviennent fréquemment. A condition que le nombre de termes dans la somme soit indépendant de  $n$ , la somme a le même ordre que

le terme d'ordre supérieur, d'après l'une des règles qui précèdent. Mais si le nombre de termes est proportionnel à la puissance de  $n$ , l'ordre de la somme dépend du nombre de termes. Le cas le plus simple est celui dans lequel  $n$  termes, chacun  $O(1)$ , sont additionnés. Le résultat est alors  $O(n)$ , à moins que les termes soient tous d'espérance nulle, et qu'un théorème de la limite centrale puisse s'appliquer (consulter la Section 4.6). Quand il en est ainsi, l'ordre de la somme est simplement  $O(n^{1/2})$ . Ainsi, si  $x_t$  est d'espérance  $\mu$  et si un théorème de la limite centrale s'y applique,

$$\sum_{t=1}^n x_t = O(n) \quad \text{et} \quad \sum_{t=1}^n (x_t - \mu) = O(n^{1/2}).$$

Utilisons maintenant l'exemple du lancé de pièce de la Section 4.2 pour illustrer l'utilisation de ces symboles d'ordre. Si nous notons  $v(b_n)$  le moment d'ordre deux de la variable aléatoire  $b_n$ , alors à partir de (4.05) nous voyons que

$$v(b_n) = \frac{1}{4}n^{-1} = O(n^{-1}).$$

De façon similaire, à partir de (4.11), nous voyons que le moment d'ordre quatre,  $E(b_n^4)$ , est la somme de deux termes, l'un  $O(n^{-2})$  et l'autre  $O(n^{-3})$ . Ainsi, nous concluons que le moment d'ordre quatre est lui-même  $O(n^{-2})$ .

Ces résultats pour  $b_n$  utilisent les relations d'ordre ordinaires, et non les relations stochastiques. Mais rappelons-nous que

$$\text{plim}_{n \rightarrow \infty} a_n = \frac{1}{2} \quad \text{et} \quad \text{plim}_{n \rightarrow \infty} b_n = 0,$$

ce qui nous permet d'écrire

$$a_n = O(1) \quad \text{et} \quad b_n = o(1).$$

Notons ici la différence. La suite  $a_n$  est du même ordre que l'unité tandis que  $b_n$  est d'un ordre inférieur. Pourtant la seule différence entre les deux est que  $b_n$  est centrée alors que  $a_n$  ne l'est pas. Les deux suites ont des limites en probabilité non stochastiques, leur variance limite étant alors nulle. Mais l'espérance limite, qui est simplement la limite en probabilité, est nulle pour  $b_n$  et non nulle pour  $a_n$ . Nous voyons donc que l'ordre d'une variable aléatoire peut dépendre des moments d'ordre un et deux de la variable (au moins). C'est précisément la soustraction du premier moment, qui apparaît dans la définition de  $b_n$ , qui nous permet de voir que le deuxième moment est d'ordre inférieur à l'unité.

Souvenons-nous à présent de l'exemple utilisé dans la Section 4.2 pour illustrer la convergence en distribution vers une limite non stochastique. Nous avons cherché une suite de variables aléatoires  $\{x_n\}$  telle que  $x_n$  était distribuée selon une  $N(0, n^{-1})$ . Notons que la variance  $n^{-1}$  est proportionnelle

à la variance de  $b_n$ , qui est  $\frac{1}{4}n^{-1}$ ; consulter (4.05). En dérivant la c.d.f. des variables aléatoires  $x_n$  nous avons vu que  $n^{1/2}x_n$  avait une distribution normale centrée réduite décrite par sa c.d.f.  $\Phi$ . De façon claire, alors, la suite  $\{x_n\}$ , multipliée par  $n^{1/2}$ , fournit une nouvelle suite à partir de laquelle la limite est une variable aléatoire distribuée suivant une  $N(0, 1)$ . Ainsi, nous avons découvert que  $n^{1/2}x_n = O(1)$ , qui implique à son tour la propriété  $x_n = o(1)$  mais qui n'est pas impliquée par elle. Cette construction nous a procuré le **taux de convergence** de la suite  $\{x_n\}$ , puisque nous pouvons maintenant affirmer que  $x_n = O(n^{-1/2})$ .

Nous allons à présent réaliser la même manœuvre avec la suite  $\{b_n\}$ , en considérant la nouvelle suite  $\{n^{1/2}b_n\}$ . A partir de (4.04) nous obtenons

$$n^{1/2}b_n = n^{-1/2} \sum_{t=1}^n z_t.$$

Evidemment,  $E(n^{1/2}b_n) = 0$ . De plus, soit directement soit à partir de (4.05), nous voyons que

$$\text{Var}(n^{1/2}b_n) = \frac{1}{4}. \quad (4.12)$$

Ainsi, nous concluons que  $n^{1/2}b_n$  est  $O(1)$ , ce qui implique que  $b_n$  elle-même est  $O(n^{-1/2})$ . Les variables aléatoires  $a_n$  et  $b_n$  sont à cet égard typiques de nombreuses variables ressemblant à des moyennes d'échantillon qui apparaissent en économétrie. La première, la moyenne de  $n$  quantités d'espérances identiques non nulles, est  $O(1)$ , alors que la dernière, la moyenne de  $n$  quantités d'espérances nulles, est  $O(n^{-1/2})$ .

Le résultat (4.12) indique de plus que si la suite  $\{n^{1/2}b_n\}$  a une limite, celle-ci sera **non dégénérée**, c'est-à-dire qu'elle ne sera pas un simple nombre non aléatoire. Nous verrons dans la Section 4.6 que cette suite possède en fait une limite, au moins en distribution, et de plus que cette limite est une variable aléatoire *normale*. C'est la propriété de normalité qui sera la conclusion d'un **théorème de la limite centrale**. De tels théorèmes, associés à des lois des grands nombres, sont fondamentaux pour toute la théorie asymptotique en statistique et en économétrie.

#### 4.4 LES PROCESSUS GÉNÉRATEURS DE DONNÉES

Dans cette section, nous appliquons la théorie mathématique développée dans les sections précédentes pour l'estimation et les tests économétriques, d'un point de vue asymptotique. Afin de préciser davantage les distributions des estimateurs et des statistiques de test, il nous faut savoir comment les données dont ils sont fonction sont générées. C'est pourquoi nous introduisons l'idée d'un processus générateur de données, ou DGP, dans la Section 2.4. Mais que signifions-nous précisément par processus générateurs de données dans un contexte asymptotique? Quand nous parlons de DGP, il suffisait de s'intéresser



à une taille d'échantillon particulière et de caractériser un DGP par une loi de probabilité qui gouverne les variables aléatoires dans un échantillon de cette taille. Mais, puisque le terme “asymptotique” fait référence à un processus limite où la taille d'échantillon tend vers l'infini, il est clair qu'une caractérisation aussi restreinte ne suffira pas. C'est pour résoudre ce problème que nous utilisons la notion de **processus stochastique**. Puisque cette notion nous permet de considérer une suite infinie de variables aléatoires, elle est parfaitement adaptée à nos besoins.

En toute généralité, un **processus stochastique** est une collection de variables aléatoires indicées par un certain ensemble d'indices convenable. Cet ensemble d'indices peut être fini, auquel cas nous n'avons pas plus qu'un vecteur de variables aléatoires. Mais cet ensemble peut aussi être infini, comportant une infinité d'éléments discrets ou continus. Nous nous intéressons presque exclusivement au cas d'une infinité discrète de variables aléatoires, de fait les suites de variables aléatoires telles que celles déjà abordées dans les sections précédentes. Pour fixer les idées, notons  $\mathbb{N}$  l'ensemble des indices, l'ensemble  $\{1, 2, \dots\}$  des nombres naturels. Alors un processus stochastique est juste l'application qui va de  $\mathbb{N}$  dans un ensemble de variables aléatoires. C'est en fait précisément ce que nous définissions précédemment par une suite de variables aléatoires, de sorte que nous comprenons que ces suites sont des cas particuliers des processus stochastiques. Elles constituent la seule sorte de processus stochastique dont nous aurons besoin dans ce livre; la notion plus générale de processus stochastique est introduite ici uniquement dans le but d'exploiter les résultats disponibles sur les processus stochastiques pour nos propos.

Le premier de ces résultats, que nous présenterons dans un moment, concerne l'existence. Le problème de l'existence est rarement un problème grave si l'on s'en tient aux mathématiques finies, mais il le devient presque toujours si le concept de l'infini apparaît. Il est assez facile de définir un ensemble fini en sélectionnant chaque objet et en décidant s'il appartient ou non à l'ensemble. Dans le cas d'un ensemble infini, une telle procédure doit être remplacée par une *règle*, comme nous l'avons remarqué précédemment dans notre discussion sur les suites. Les règles ont beau exister, il n'est pas du tout évident qu'une règle donnée définisse quelque chose — d'où la question d'existence — et, si c'est le cas, que ce qu'elle définit soit intéressant.

Pour les processus stochastiques en général, l'existence est établie par un célèbre théorème dû à l'éminent mathématicien, probabiliste et physicien soviétique, A. N. Kolmogorov. Comme il a marqué de son empreinte de nombreux domaines de la recherche, les lecteurs peuvent avoir déjà rencontré son nom auparavant. Le contenu de son théorème est qu'une suite de variables aléatoires est bien définie par une règle si la règle génère pour chaque sous-suite *finie* une distribution de probabilité jointe de dimension finie *compatible*, dans un certain sens, avec celles générées pour toutes les autres sous-suites finies. Ainsi, il n'est jamais nécessaire de considérer des distributions de di-

mension infinie, même en supposant qu'une manière pourrait être trouvée pour procéder de la sorte. En effet, le **théorème d'existence de Kolmogorov** nous enseigne que si deux conditions de compatibilité sont satisfaites, une suite aléatoire bien définie existe.

Ces deux conditions de compatibilité sont très intuitives. La première nécessite que si l'on demande (de la règle) la distribution des variables aléatoires indicées par un ensemble fini d'indices,  $\{t_1, t_2, \dots, t_n\}$ , par exemple, et si après l'on demande la distribution des variables aléatoires indicées par une permutation ou un mélange des indices de cet ensemble, alors on doit recevoir la bonne réponse. La seconde condition nécessite que si l'on demande la fonction de distribution des variables indicées par un ensemble fini quelconque  $\{t_1, t_2, \dots, t_n\}$ , et à nouveau la distribution des variables indicées par un *sous-ensemble* de  $\{t_1, t_2, \dots, t_n\}$ , alors la réponse doit être la distribution marginale qui serait calculée de la manière standard à partir de la fonction de distribution de l'ensemble des variables.

Ces deux conditions simples et facilement vérifiables sont suffisantes pour régler le problème de l'existence. Nous considérons alors une suite de variables aléatoires potentiellement vectorielles que nous pouvons noter à la manière habituelle  $\{\mathbf{y}_t\}_{t=1}^\infty$ . Pour que cette suite soit un processus stochastique bien défini, d'après le théorème d'existence de Kolmogorov, il suffit d'être capable de définir la distribution de probabilité jointe de n'importe quel sous-ensemble fini des éléments  $\mathbf{y}_t$  de la suite d'une manière compatible avec toutes les autres distributions de probabilité jointes des sous-ensembles finis des  $\mathbf{y}_t$ . Il s'avère que ceci est équivalent à deux conditions simples. Tout d'abord, nous devons être capables de définir la distribution de probabilité jointe des variables aléatoires appartenant à n'importe quel **échantillon** de taille finie, c'est-à-dire un sous-ensemble de la forme  $\mathbf{y}^n \equiv \{\mathbf{y}_t\}_{t=1}^n$ , pour une **taille d'échantillon**  $n$  finie quelconque.<sup>2</sup> Ensuite, la distribution de l'échantillon de taille  $n$  devrait être la distribution obtenue par l'intégration de la distribution de l'échantillon de taille  $n + 1$  par rapport la dernière variable. Puisque nous souhaiterions satisfaire ces conditions dans toutes les circonstances concevables, nous avons la chance que l'idée mathématique d'un processus stochastique ou d'une suite aléatoire corresponde exactement à ce qui est nécessaire pour construire une théorie asymptotique en économétrie.

Pour parler de théorie asymptotique, nous devons pouvoir définir un DGP. Pour cela, nous devons être capables de spécifier la distribution jointe de l'ensemble des variables aléatoires correspondant aux observations contenues dans un échantillon de taille arbitrairement grande. Il s'agit à l'évidence d'une condition très forte. En économétrie, ou dans toute autre discipline, nous traitons exclusivement des échantillons finis. Comment pouvons-nous alors,

<sup>2</sup> Nous utiliserons cette sorte de notation assez généralement. Un indice identifiera une observation particulière, tandis qu'un exposant identifiera la taille d'échantillon.

même de façon théorique, traiter des échantillons infinis? La réponse est la même que celle qui nous a permis de traiter des suites infinies en général: nous devons d'une certaine façon créer une *règle* qui échantillonne finis un processus stochastique infini. Malheureusement, pour n'importe quel contexte d'observation, il existe un nombre infini de manières qui permettent de construire une telle règle, et des règles différentes peuvent conduire à des conclusions asymptotiques très divergentes.

L'exemple simple suivant illustre les concepts en cause. Il est fréquent dans la pratique d'associer une **tendance temporelle**  $\boldsymbol{\tau} \equiv [1 : 2 : 3 \cdots n]$  aux régresseurs d'un modèle de régression linéaire. La tendance temporelle peut aussi être définie par la définition de chacune de ses composantes:

$$\tau_t = t, \quad t = 1, 2, 3 \cdots. \quad (4.13)$$

L'exemple est un modèle dans lequel cette tendance temporelle est le seul régresseur:

$$\mathbf{y} = \alpha \boldsymbol{\tau} + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}. \quad (4.14)$$

L'estimation du paramètre  $\alpha$  est alors

$$\hat{\alpha} = (\boldsymbol{\tau}^\top \boldsymbol{\tau})^{-1} \boldsymbol{\tau}^\top \mathbf{y}. \quad (4.15)$$

En supposant que le DGP est en effet (4.14) avec  $\alpha = \alpha_0$  et  $\sigma^2 = \sigma_0^2$ , il est facile de voir que cette estimation devient

$$\hat{\alpha} = \alpha_0 + (\boldsymbol{\tau}^\top \boldsymbol{\tau})^{-1} \boldsymbol{\tau}^\top \mathbf{u}. \quad (4.16)$$

Evidemment, les propriétés de l'estimation (4.15) dépendront des propriétés du terme aléatoire  $(\boldsymbol{\tau}^\top \boldsymbol{\tau})^{-1} \boldsymbol{\tau}^\top \mathbf{u}$ . Si la taille de l'échantillon est  $n$ ,

$$\boldsymbol{\tau}^\top \boldsymbol{\tau} = \sum_{t=1}^n t^2 = \frac{1}{6} n(2n+1)(n+1) = O(n^3).$$

On montre aisément que l'expression utilisée ici pour la somme  $\sum_{t=1}^n t^2$  est correcte par induction. Le second facteur dans l'expression (4.16),  $\boldsymbol{\tau}^\top \mathbf{u}$ , sera d'espérance nulle comme d'habitude (ce qui implique que  $\hat{\alpha}$  est certainement sans biais) et de variance  $\sigma_0^2 \boldsymbol{\tau}^\top \boldsymbol{\tau}$ . Ainsi, la variance de  $\hat{\alpha}$  est

$$\text{Var}(\hat{\alpha}) = \sigma_0^2 (\boldsymbol{\tau}^\top \boldsymbol{\tau})^{-1} = O(n^{-3}). \quad (4.17)$$

Dans cet exemple, il est naturel de développer la tendance  $\boldsymbol{\tau}$  à des tailles d'échantillon arbitrairement grandes au moyen de la règle (4.13), et c'est en procédant de la sorte que nous avons obtenu l'ordre de la variance de  $\hat{\alpha}$ ,  $O(n^{-3})$ . Mais ce *n'est pas* le genre de règle toujours employée dans l'analyse asymptotique. L'hypothèse plus fréquemment formulée, comme nous l'avons

fait dans la Section 3.5 en connexion avec les propriétés des statistiques de test, est que les régresseurs sont fixes dans des échantillons répétés. Dans le contexte actuel, ceci signifie que la règle utilisée pour développer un échantillon fini en un échantillon arbitrairement grand est la suivante: pour une taille d'échantillon observée  $m$ , on considère *seulement* des échantillons de taille  $n \equiv Nm$ , pour des entiers positifs  $N$ . Clairement, quand  $N \rightarrow \infty$  la taille d'échantillon  $n$  tend également vers l'infini. Puis dans chaque *répétition* de l'échantillon (d'où la terminologie "échantillons répétés") les régresseurs sont supposés être les *mêmes* que ceux de l'échantillon observé, et seules des perturbations aléatoires sont supposées être différentes. De fait, pour n'importe quelle taille d'échantillon  $n$ , on suppose habituellement que  $\mathbf{u}$  est un vecteur de dimension  $n$  tel que  $\mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$  où  $\mathbf{I}_n$  est la matrice identité de dimension  $n$  et  $\sigma_0^2$  est une variance constante quelconque.

Cette idée de régresseurs fixes en échantillons répétés est naturellement excellente pour les **données en coupe transversale**, c'est-à-dire les ensembles de données dans lesquels les observations sont toutes enregistrées à la même période. Ces observations seront typiquement celles d'une population supposée caractérisée par certaines régularités statistiques que l'on désire estimer, ou par l'existence de caractéristiques que l'on désire tester. Dans ce cas, l'idée de régresseurs fixes en échantillons répétés correspond à l'hypothèse que, en collectant davantage de données, elles correspondront à "davantage des mêmes choses." Pour exprimer cette idée différemment, cela équivaut à supposer que l'échantillon collecté est suffisamment *représentatif* de la population étudiée. La plupart des ensembles de données utilisés en micro-économétrie sont des ensembles en coupe transversale. Ils seront souvent constitués de données sur les agents économiques individuels tels les firmes ou les ménages, comme dans l'exemple du propriétaire de son domicile de la Section 4.2.

Les difficultés abordées dans cette section sur les manières de développer des DGPs à des échantillons arbitrairement grands ne concernent généralement pas les données en coupe transversale.<sup>3</sup> Elles concernent au contraire les **données temporelles**. Pour de tels ensembles de données, les observations sont ordonnées de manière chronologique; à chaque observation correspond une *date*, peut-être simplement une année, mais peut-être un trimestre, un mois, un jour, ou même, dans le cas de certains ensembles de données sur les marchés financiers, une heure ou une minute. Les difficultés du développement de l'échantillon surviennent même avec plus d'insistance dans le cas de ce que l'on appelle **données de panel**; consulter la Section 9.10. De telles données sont obtenues quand une coupe transversale est observée à plusieurs moments différents. Ainsi, le même sujet de la population étudiée, habituellement soit un ménage soit une firme, peut être suivi dans le temps. Pour réaliser une

<sup>3</sup> Cet exposé peut être trop optimiste. Considérons, par exemple, une coupe transversale sur des pays. Si nous avons un ensemble de données contenant de l'information sur tous les pays de OCDE, il est difficile d'imaginer une quelconque règle pour développer l'échantillon!

analyse asymptotique sur de telles données, il est nécessaire de définir une règle pour développer la taille de l'échantillon à l'infini de deux manières différentes, l'une correspondant à la dimension transversale des données et l'autre à leur dimension temporelle. Dans notre discussion ultérieure sur les difficultés liées au développement des DGP à de grands échantillons, nous discuterons toujours, explicitement ou pas, de données temporelles ou de la dimension temporelle des données de panel.

Posons l'hypothèse de régresseurs fixes en échantillons répétés pour le modèle (4.14). Alors, si la taille d'échantillon observé était  $m$ , pour une taille d'échantillon  $n = Nm$ , nous devrions avoir

$$\hat{\alpha} = \alpha_0 + \left( N(\boldsymbol{\tau}^m)^\top \boldsymbol{\tau}^m \right)^{-1} \sum_{k=1}^N (\boldsymbol{\tau}^m)^\top \mathbf{u}_k,$$

où  $\boldsymbol{\tau}^m$  désigne une tendance comportant seulement  $m$  éléments, et où les  $\mathbf{u}_k$ ,  $k = 1, 2, \dots, N$ , sont des vecteurs aléatoires indépendants de dimension  $m$  ayant les propriétés habituelles. Alors

$$\begin{aligned} N(\boldsymbol{\tau}^m)^\top \boldsymbol{\tau}^m &= \frac{1}{6}n(2m+1)(m+1) = O(n) \quad \text{et} \\ \text{Var} \left( \sum_{k=1}^N (\boldsymbol{\tau}^m)^\top \mathbf{u}_k \right) &= \frac{1}{6}\sigma_0^2 n(2m+1)(m+1) = O(n), \end{aligned}$$

ce qui implique que

$$\text{Var}(\hat{\alpha}) = \sigma_0^2 \left( \frac{1}{6}n(2m+1)(m+1) \right)^{-1} = O(n^{-1}). \quad (4.18)$$

Une comparaison entre (4.17) et (4.18) révèle que le comportement de l'estimateur  $\hat{\alpha}$  est très différent selon les deux règles différentes de développement de la taille d'échantillon.

Il n'existe pas toujours de résolution simple au type de problème posé dans l'exemple précédent. Il n'est *habituellement* pas réaliste de supposer que les tendances temporelles linéaires de la forme de  $\boldsymbol{\tau}$  seront toujours continûment croissantes, mais il suffit d'examiner les séries de prix pendant le vingtième siècle (ou de nombreux autres siècles) pour réaliser que certaines variables économiques ne semblent pas posséder de borne supérieure naturelle. Même les séries en quantité telles le PIB réel ou la consommation par tête sont parfois considérées comme non bornées. Néanmoins, bien que les théories asymptotiques qui résultent de différentes sortes de règles pour le développement des DGP à des échantillons arbitrairement grands puissent être très différentes, il est important que soit clair le fait que la sélection d'une théorie asymptotique parmi toutes celles en compétition *ne constitue pas* un problème empirique. Pour n'importe quelle question empirique donnée, l'échantillon est ce qu'il est, même si la *possibilité* de collecter davantage de

données pertinentes existe. Le problème consiste toujours à sélectionner un *modèle* convenable, non seulement pour que les données existent, mais aussi pour qu'existe un ensemble de *phénomènes* économiques dont les données sont supposées être des manifestations. Il existe toujours une infinité de modèles (pas tous plausibles naturellement) compatibles avec n'importe quel ensemble de données. En conséquence, le problème de sélection d'un modèle parmi un ensemble de tels modèles peut ne trouver de solution que sur la base de critères tels que le pouvoir explicatif des concepts utilisés dans le modèle, la simplicité de son expression, ou la facilité d'interprétation, et non sur la base des informations contenues dans les données elles-mêmes.

Bien que, dans le modèle (4.14), l'hypothèse que la variable de tendance temporelle tende vers l'infini avec la taille d'échantillon puisse sembler plus plausible que l'hypothèse de régresseurs fixes en échantillons répétés, nous supposons dans la majeure partie de ce livre que le DGP est plutôt du dernier type que du premier. Le problème de laisser croître  $t_t$  vers l'infini avec la taille d'échantillon est que chaque observation supplémentaire nous donne davantage d'information sur la valeur de  $\alpha$  que n'importe quelle autre observation précédente. C'est pourquoi  $\text{Var}(\hat{\alpha})$  s'avérait être  $O(n^{-3})$  quand nous avons posé cette hypothèse concernant le DGP. Il semble beaucoup plus plausible dans la plupart des cas que chaque observation supplémentaire donne, en moyenne, la même quantité d'information que les observations précédentes. Ceci implique que la variance des estimations paramétriques sera  $O(n^{-1})$ , comme l'était  $\text{Var}(\hat{\alpha})$  quand nous supposions que le DGP était composé de régresseurs fixes en échantillons répétés. Nos hypothèses générales concernant les DGP mèneront de plus à la conclusion que la variance des estimations paramétriques est  $O(n^{-1})$ , bien que nous considérerons des DGP qui ne conduisent pas à cette conclusion dans le Chapitre 20, à propos des modèles dynamiques.

## 4.5 CONVERGENCE ET LOIS DES GRANDS NOMBRES

Nous commençons cette section en introduisant la notion de **convergence**, une des idées les plus fondamentales de la théorie asymptotique. Quand nous nous intéressons à l'estimation des paramètres à partir des données, il est souhaitable que les estimations paramétriques possèdent certaines propriétés. Dans les Chapitres 2 et 3, nous avons vu que, sous certaines conditions de régularité, l'estimateur OLS est sans biais et suit une distribution normale avec une matrice de covariance connue à un facteur multiplicatif de la variance des aléas près, dont on peut avoir une estimation sans biais. Nous n'étions pas capables dans ces chapitres de démontrer un quelconque résultat correspondant pour l'estimateur NLS, et nous remarquons justement que la théorie asymptotique serait nécessaire. La convergence est la première des **propriétés asymptotiques** souhaitables pour un estimateur. Dans le Chapitre

5 nous fournirons des conditions sous lesquelles l'estimateur NLS est convergent. Ici nous nous contenterons d'introduire la notion et d'illustrer le lien étroit qui existe entre les lois des grands nombres et les démonstrations de la convergence.

Un estimateur  $\hat{\beta}$  d'un vecteur de paramètres  $\beta$  est dit **convergent** s'il converge vers sa vraie valeur quand la taille de l'échantillon tend vers l'infini. Cet exposé n'est pas faux ni même sérieusement trompeur, mais il repose sur un certain nombre d'hypothèses implicites et utilise des termes non définis. Essayons de rectifier ceci et par là même d'acquiescer une meilleure compréhension de ce qu'est la convergence.

Tout d'abord, comment un *estimateur* peut-il converger? Il le peut si nous le convertissons en une suite. Pour cela, nous écrivons  $\hat{\beta}^n$  pour l'estimateur obtenu à partir d'un échantillon de taille  $n$  et définissons ensuite l'estimateur  $\hat{\beta}$  comme la suite  $\{\hat{\beta}^n\}_{n=m}^{\infty}$ . La limite inférieure  $m$  de la suite sera habituellement supposée être la taille d'échantillon la plus petite qui permet de calculer  $\hat{\beta}^n$ . Par exemple, si  $\mathbf{y}^n$  désigne la régressande et  $\mathbf{X}^n$  la matrice des régresseurs pour une régression linéaire sur un échantillon de taille  $n$ , et si  $\mathbf{X}^n$  est une matrice de dimension  $n \times k$ , alors  $m$  ne peut pas être inférieur à  $k$ , le nombre de régresseurs. Pour  $n > k$  nous avons comme d'habitude l'estimation  $\hat{\beta}^n = ((\mathbf{X}^n)^\top \mathbf{X}^n)^{-1} (\mathbf{X}^n)^\top \mathbf{y}^n$ , et cette formule contient la règle qui génère la suite  $\hat{\beta}$ .

Un élément d'une suite  $\hat{\beta}$  est une variable aléatoire. S'il doit converger vers une véritable valeur quelconque, il faut préciser de quelle sorte de convergence nous parlons, puisque nous avons vu que plusieurs sortes de convergences sont disponibles. Si nous utilisons la convergence presque sûre, nous dirons que nous avons une **convergence forte** ou que l'estimateur est **fortement convergent**. Parfois une telle propriété existe. Plus fréquemment, nous utilisons la convergence en probabilité, et obtenons ainsi seulement la **convergence faible**. Les termes "forte" et "faible" sont utilisés dans le même sens que dans les définitions des lois forte et faible des grands nombres.

Ensuite, qu'entendons-nous par "vraie valeur"? Nous répondons à cette question en détail dans le prochain chapitre, mais nous devons ici au moins noter que la convergence d'une suite de variables aléatoires vers n'importe quelle sorte de limite dépend de la règle, ou du DGP, qui a généré la suite. Par exemple, si la règle assure que, pour n'importe quelle taille d'échantillon  $n$ , la régressande et la matrice des régresseurs d'une régression linéaire sont reliées par l'équation

$$\mathbf{y}^n = \mathbf{X}^n \beta_0 + \mathbf{u}^n, \quad (4.19)$$

pour un vecteur  $\beta_0$  fixe quelconque, où  $\mathbf{u}^n$  est un vecteur de dimension  $n$  d'aléas bruits blancs, alors la véritable valeur pour ce DGP sera  $\beta_0$ . L'estimateur  $\hat{\beta}$ , pour être utile, devra converger vers  $\beta_0$  sous le DGP (4.19) *quelle que soit* la valeur fixée  $\beta_0$ . Cependant, si le DGP est tel que (4.19) n'est pas valable pour un quelconque  $\beta_0$ , alors nous ne trouvons aucune signification au terme "convergence" tel que nous l'utilisons à présent.

Après ce préambule, nous pouvons finalement étudier la convergence dans un cas particulier. Nous pourrions prendre comme exemple la régression linéaire (4.19), mais cela nous mènerait à gérer trop de problèmes connexes dont nous discuterons dans le prochain chapitre. Au lieu de cela, nous considérerons l'exemple très instructif que procure le **Théorème Fondamental de la Statistique**, une version simple de ce que nous allons démontrer. Ce théorème, qui est en effet fondamental pour toute inférence statistique, établit que si nous échantillonnons de façon aléatoire avec remise à partir d'une population, la fonction de distribution empirique converge vers la fonction de distribution de la population.

Formalisons cet exposé et démontrons-le. Le terme **population** est utilisé dans son sens statistique d'ensemble, fini ou infini, à partir duquel des **tirages aléatoires** sont opérés. Chaque tirage est un membre de la population. Nous entendons par **échantillonnage aléatoire avec remise** une procédure qui garantit la constance de la probabilité que chaque membre de la population soit tiré à chaque tirage. Un **échantillon aléatoire** sera un ensemble fini de tirages. Formellement, la population est représentée par une c.d.f.  $F(x)$  pour une variable aléatoire scalaire  $x$ . Les tirages dans la population sont caractérisés par des valeurs différentes et indépendantes de  $x$ .

Un échantillon aléatoire de taille  $n$  peut être noté par  $\{Y_t\}_{t=1}^n$ , où les  $Y_t$  sont des réalisations indépendantes. Alors, nous entendons par **fonction de distribution empirique** générée par l'échantillon, la fonction de distribution suivante

$$\hat{F}^n(x) \equiv \frac{1}{n} \sum_{t=1}^n I_{(-\infty, x)}(Y_t). \quad (4.20)$$

La **fonction indicatrice**  $I$  associée à l'intervalle  $(-\infty, x)$  prend la valeur 1 si son argument appartient à l'intervalle et 0 sinon. (Des fonctions indicatrices peuvent être définies de façon similaire pour n'importe quel sous-ensemble de la droite réelle ou de n'importe quel autre espace sur lequel les variables aléatoires prennent leurs valeurs.) Nous laissons comme exercice la démonstration que l'expression (4.20) définit en effet la fonction de distribution pour la distribution discrète qui attribue une masse de probabilité  $n^{-1}$  à chaque réalisation contenue dans l'échantillon  $\{Y_t\}_{t=1}^n$ .

Ensuite, nous passons de la fonction de distribution empirique (4.20), associée à un échantillon donné, à une fonction de distribution *aléatoire*. A cette fin, les réalisations  $Y_t$  sont remplacées par les variables aléatoires  $y_t$ . Comme nous supposons que les différents tirages d'un échantillon réel étaient indépendants, nous supposons à présent que les différentes variables aléatoires  $y_t$  sont indépendantes. De fait, nous traitons d'un DGP qui peut générer des suites aléatoires  $\{y_t\}_{t=1}^n$  de longueur arbitraire  $n$ . Pour n'importe quel  $n$  donné, nous avons alors une c.d.f. aléatoire telle que:

$$\hat{F}^n(x) = \frac{1}{n} \sum_{t=1}^n I_{(-\infty, x)}(y_t). \quad (4.21)$$



La fonction de distribution empirique d'un échantillon aléatoire  $\{Y_t\}_{t=1}^n$  est par conséquent une réalisation de (4.21). Pour démontrer le Théorème Fondamental de la Statistique, nous devons montrer que, pour tout réel  $x$ ,  $\hat{F}^n(x)$  tend en probabilité vers  $F(x)$  quand  $n \rightarrow \infty$ .

Tout d'abord, fixons-nous une valeur réelle quelconque de  $x$ . Nous pouvons alors observer que chaque terme dans la somme dans le membre de droite de (4.21) ne dépend que de la variable  $y_t$ . Puisque les  $y_t$  sont mutuellement indépendants ces termes le sont par conséquent aussi. Chaque terme utilise la *même* valeur fixe de  $x$ , de sorte que dans chaque terme nous avons la *même* fonction. Puisque les  $y_t$  suivent tous la même distribution, alors c'est aussi le cas des termes de la somme. Par conséquent (4.21) est la moyenne de  $n$  termes aléatoires, tous mutuellement indépendants et suivant tous la même distribution. Cette distribution n'est pas difficile à décrire. Une fonction indicatrice, par construction, peut prendre seulement une des deux valeurs, 0 et 1. Nous aurons alors décrit complètement la distribution si nous donnons la probabilité associée à chacune de ces deux valeurs. D'après la définition de  $I$ ,

$$\begin{aligned} \Pr(I_{(-\infty, x)}(y_t) = 1) &= \Pr(y_t \in (-\infty, x)) \\ &= \Pr(y_t < x) \\ &= F(x), \end{aligned} \tag{4.22}$$

où la dernière ligne provient de la définition de la c.d.f.  $F(\cdot)$ . La probabilité complémentaire,  $\Pr(I_{(-\infty, x)}(y_t) = 0)$ , est alors naturellement  $1 - F(x)$ .

Dans la Section 4.2, nous avons démontré une loi faible des grands nombres pour l'exemple du lancé de pièce. Nous avons cherché la moyenne d'une suite de variables aléatoires indépendantes et identiquement distribuées avec une distribution similaire à celle des fonctions indicatrices, pour lesquelles seulement deux valeurs, 0 et 1, sont possibles. A cause de l'hypothèse d'absence de biais de la pièce, la contrainte que chaque valeur soit associée à une probabilité d'un demi était ajoutée. Le problème est ici totalement identique au problème de la proportion limite des faces dans une suite de lancers d'une pièce *biaisée*. Il serait simple de modifier la démonstration de Section 4.2 afin de la rendre applicable à ce cas. Au lieu de cela, nous préférons démontrer maintenant une loi générale (faible) des grands nombres, qui comprendra ce cas ainsi que de nombreux autres.

*Théorème 4.1. Loi Faible Simple des Grands Nombres. (Chebyshev)*

Supposons que pour chaque entier positif  $n$  nous avons un ensemble de variables aléatoires scalaires indépendantes  $\{y_1^n, y_2^n, \dots, y_n^n\}$ . Soit

$$S_n \equiv \sum_{i=1}^n y_i^n,$$

et supposons de plus que  $E(y_i^n) = m_i^n$  et que  $\text{Var}(y_i^n) = v_i^n$ . Alors la limite en probabilité quand  $n \rightarrow \infty$  de la suite

$$\left\{ N_n^{-1} \left( S_n - \sum_{i=1}^n m_i^n \right) \right\}$$

est zéro pour toutes les suites non aléatoires des nombres réels positifs  $N_n$  qui tendent vers l'infini avec  $n$  et tels que

$$V_n \equiv \sum_{i=1}^n v_i^n = o(N_n^2).$$

*Démonstration:* la technique de démonstration est la même que pour la loi faible démontrée dans la Section 4.2 pour l'exemple du lancé de pièce: nous exploitons l'inégalité de Chebyshev. Notons tout d'abord que

$$\text{Var} \left( N_n^{-1} \left( S_n - \sum_{i=1}^n m_i^n \right) \right) = N_n^{-2} V_n \equiv w_n.$$

Il est clair que quand  $n \rightarrow \infty$ ,  $w_n \rightarrow 0$ , parce que  $V_n = o(N_n^2)$ , et quand  $n$  tend vers l'infini,  $N_n$  en fait de même. Mais nous avons observé dans la Section 4.2 que n'importe quelle suite centrée de variables aléatoires pour lesquelles les variances tendent vers zéro quand  $n$  tend vers l'infini tend en probabilité vers zéro. Ainsi la démonstration est complète.

Notons que ce théorème nécessite l'existence à la fois du premier et du second moment des variables  $y_i^n$ . Dans la Section 4.7, nous listerons, sans les démontrer, des théorèmes basés sur des conditions de régularité moins strictes. Par ailleurs, nous avons introduit un degré de généralité raisonnable. Nos variables aléatoires  $y_i^n$  doivent encore être indépendantes (cette contrainte sera relâchée dans la Section 4.7), mais elles peuvent posséder des espérances et des variances différentes, et naturellement des moments d'ordre supérieur différents, à condition qu'ils existent. De plus, la dépendance explicite de  $y_i^n$  à  $n$  signifie que, quand une taille d'échantillon croît, les variables aléatoires indicées par des valeurs faibles de  $n$  ne sont pas inexorablement les mêmes que ce qu'elles étaient dans les échantillons précédents, plus petits.

Clairement, le Théorème 4.1 fournit plus que nécessaire pour conclure que

$$\left\{ \frac{1}{n} \sum_{t=1}^n I_{(-\infty, x)}(y_t) - F(x) \right\}$$

tend en probabilité vers zéro quand  $n \rightarrow \infty$ . De (4.22), nous trouvons que

$$\begin{aligned} E(I_{(-\infty, x)}(y_t)) &= F(x) \quad \text{et} \\ \text{Var}(I_{(-\infty, x)}(y_t)) &= F(x)(1 - F(x)). \end{aligned}$$

Ainsi, la variance  $V_n$  dans ce cas est  $nF(x)(1-F(x))$ . Cette dernière quantité est  $O(n)$ , et par conséquent aussi  $o(n^2)$ , comme le demande le théorème avec  $N_n = n$ . Nous avons alors terminé la démonstration de notre version du Théorème Fondamental de la Statistique.

Cette démonstration peut servir de modèle pour les démonstrations de convergence de nombreux estimateurs simples. Par exemple, nous désirons souvent estimer les **moments** d'une certaine distribution. Si  $x$  est une variable aléatoire obéissant à cette distribution, alors par définition le moment d'ordre  $k$  de la distribution est  $\mu_k \equiv E(x^k)$ . L'estimateur le plus intuitif pour  $\mu_k$  est le **moment d'échantillon** correspondant. Le moment d'échantillon d'ordre  $k$   $M_k$  issu d'un échantillon observé  $\{Y_t\}_{t=1}^n$  est défini comme le moment d'ordre  $k$  de la distribution empirique de l'échantillon, qui est bien évidemment

$$M_k = \frac{1}{n} \sum_{t=1}^n Y_t^k.$$

Si comme auparavant, nous remplaçons les observations  $Y_t$  dans la somme précédente par des variables aléatoires  $y_t$ , et le moment réalisé  $M_k$  par une variable aléatoire  $m_k$ , nous constatons que  $m_k$  est une estimation convergente de  $\mu_k$  en considérant  $m_k - \mu_k$ , à savoir

$$m_k - \mu_k = \frac{1}{n} \sum_{t=1}^n (y_t^k - E(y_t^k)).$$

Les variables aléatoires  $y_t^k - E(y_t^k)$  sont centrées, de sorte que le membre de droite de cette équation satisfera les conditions du Théorème 4.1 si le moment de la population d'ordre  $(2k)$  existe. Ainsi, sous cette condition assez faible, nous concluons que les moments d'échantillon estiment de façon convergente les moments de la population.

Tout comme il est presque obligatoire en échantillon fini de démontrer qu'un estimateur est sans biais, il est presque obligatoire dans l'analyse asymptotique de démontrer qu'un estimateur converge. Bien que l'on puisse inférer à partir de ceci que la convergence est l'équivalent asymptotique d'une **absence de biais**, cela ne serait pas correct. La convergence n'implique pas, et n'est pas impliquée par l'absence de biais. Un estimateur peut être sans biais et convergent, biaisé mais convergent, sans biais mais non convergent, ou biaisé et non convergent! Considérons les exemples suivants, qui traitent tous de l'estimation de la moyenne d'une population caractérisée par une espérance  $\mu$  et une variance  $\sigma^2$ , basée sur un échantillon de  $n$  observations  $y_t$ . Nous avons déjà vu que la moyenne d'échantillon  $m_1$  était convergente et sans biais. Mais considérons maintenant les estimateurs

$$\tilde{\mu} = \frac{1}{n-3} \sum_{t=1}^n y_t, \quad (4.23)$$

$$\ddot{\mu} = \frac{1}{2}y_1 + \frac{1}{2(n-1)} \sum_{t=2}^n y_t, \text{ et} \quad (4.24)$$

$$\tilde{\mu} = \frac{2}{n} \sum_{t=1}^n y_t. \quad (4.25)$$

Le premier de ceux-ci, (4.23), est à l'évidence *biaisé*, parce que

$$E(\tilde{\mu}) = \frac{1}{n-3} \sum_{t=1}^n \mu = \frac{n}{n-3} \mu,$$

mais il est néanmoins *convergent*, parce que  $\tilde{\mu} = (n/(n-3))m_1$  et  $n/(n-3) \rightarrow 1$  quand  $n \rightarrow \infty$ . D'un autre côté, le deuxième estimateur, (4.24), est *sans biais*, parce que

$$E(\ddot{\mu}) = \frac{1}{2}\mu + \frac{1}{2(n-1)} \sum_{t=2}^n \mu = \mu.$$

Cependant,  $\ddot{\mu}$  est *non convergent*, parce qu'il a une variance qui ne tend pas vers 0 quand  $n \rightarrow \infty$ , puisque  $y_1$  prend toujours un poids fini quelle que soit la taille de l'échantillon. Ceci signifie que  $\ddot{\mu}$  ne peut certainement pas avoir une limite en probabilité non stochastique, contrairement à tout estimateur convergent. Finalement, le troisième estimateur, (4.25), est nettement biaisé, et il est non convergent puisqu'il convergera vers  $2\mu$  à la place de  $\mu$ .

La relation entre la convergence et l'**absence de biais asymptotique** est encore plus subtile, parce qu'à première vue il existe deux définitions possibles pour le dernier concept. La première est qu'un estimateur est asymptotiquement sans biais si l'espérance de sa distribution asymptotique est la véritable valeur du paramètre. La seconde est qu'un estimateur est asymptotiquement sans biais si la limite des espérances des variables aléatoires qui constituent l'estimateur (une suite) est la véritable valeur du paramètre. Ces deux définitions ne sont pas équivalentes. La raison technique de cette divergence est que le calcul d'un moment d'une variable aléatoire n'induit une application continue de l'ensemble des variables aléatoires dans  $\mathbb{R}$  dans aucune des topologies induites sur les ensembles de variables aléatoires par les différentes sortes de convergence stochastiques considérées.

Pour repérer la source du problème, considérons l'exemple "pathologique" suivant. Voici un estimateur manifestement étrange du paramètre scalaire  $\theta$ :

$$\hat{\theta}^n \equiv \begin{cases} \theta & \text{avec la probabilité } 1 - n^{-1} \\ 2n\theta & \text{avec la probabilité } n^{-1}. \end{cases}$$

Cet estimateur est évidemment convergent: pour n'importe quel  $\varepsilon > 0$ ,

$$\Pr(|\hat{\theta}^n - \theta| > \varepsilon) \leq \Pr(\hat{\theta}^n = 2n\theta) = n^{-1},$$

qui est inférieure à  $\delta$  pour tout  $n > \delta^{-1}$ . L'espérance de cet estimateur existe:

$$E(\hat{\theta}^n) = 3\theta - n^{-1}\theta,$$

et cette espérance tend vers une limite bien définie de  $3\theta$  quand  $n \rightarrow \infty$ . La limite de l'espérance n'est pas par conséquent la véritable valeur du paramètre. Cependant, la distribution asymptotique de l'estimateur, à savoir la distribution de la variable aléatoire limite vers laquelle tend la suite  $\{\hat{\theta}^n\}$ , a une espérance assez différente, et correcte, égale à  $\theta$ . Un exercice utile consiste à écrire la c.d.f. de  $\hat{\theta}^n$  et à montrer que celle-ci converge en chaque point vers une distribution dégénérée concentrée en  $\theta$ .

Il est clair à partir de cet exemple que la définition que nous voudrions pour la distribution asymptotique sans biais est celle utilisant la limite des espérances, et non l'espérance limite, puisque cette dernière est une propriété de la variable aléatoire que l'on n'observe jamais dans un monde fini. Les concepts asymptotiques ne fournissent pas des approximations utiles en présence de discontinuités en l'infini! Avec cette définition, cependant, la convergence n'implique pas l'absence de biais asymptotique, à moins d'écarter les exemples pathologiques comme celui-ci. De tels exemples peuvent souvent être éliminés, naturellement. Dans cet exemple particulier, la *variance* de  $\hat{\theta}^n$  est  $O(n)$  quand  $n \rightarrow \infty$  et par conséquent, elle ne tend pas vers une limite finie. Cette pathologie ne se manifeste pas avec un estimateur convergent, tel  $m_k$  pour  $\mu_k$  défini auparavant, qui satisfait une loi des grands nombres telle que le Théorème 4.1. Dans un cas semblable, le centrage de la suite opéré pour des tailles d'échantillon *finies* garantit que l'espérance de la distribution limite et la limite des espérances sont identiques. Un problème exactement similaire surviendra dans la prochaine section dans le contexte des théorèmes de la limite centrale et des variances limites.

## 4.6 THÉORÈMES DE LA LIMITE CENTRALE

Il existe la même sorte de relation étroite entre la propriété de **normalité asymptotique** et les théorèmes de la limite centrale qu'entre la convergence et la loi des grands nombres. La manière la plus simple de démontrer cette relation étroite consiste à travailler sur un exemple. Supposons que des échantillons soient générés par des tirages aléatoires issus de distributions d'espérance inconnue  $\mu$  et de variances variables inconnues. Par exemple, il se pourrait que la variance de la distribution à partir de laquelle l'observation  $t$  est tirée soit

$$\sigma_t^2 \equiv \omega^2 \left(1 + \frac{1}{2}(t \bmod 3)\right). \quad (4.26)$$

Alors,  $\sigma_t^2$  prendra les valeurs  $\omega^2$ ,  $1.5\omega^2$ , et  $2\omega^2$  avec des probabilités identiques. Ainsi,  $\sigma_t^2$  varie systématiquement avec  $t$  mais toujours dans certaines limites, dans ce cas  $\omega^2$  et  $2\omega^2$ .

Nous supposons que l'utilisateur ne connaît pas la relation exacte (4.26) et pose l'hypothèse que les variances  $\sigma_t^2$  varient entre deux bornes positives et ont une moyenne asymptotique d'une valeur quelconque  $\sigma_0^2$ , connue ou pas, définie par

$$\sigma_0^2 \equiv \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \sigma_t^2 \right).$$

La moyenne d'échantillon peut à nouveau être utilisée comme un estimateur de la moyenne de la population, puisque notre loi des grands nombres, le Théorème 4.1, s'applique. L'utilisateur est également prêt à supposer que les distributions dont les observations sont issues ont des moments d'ordre trois bornés, aussi le supposons-nous également. L'utilisateur souhaite réaliser une inférence statistique asymptotique sur l'estimation issue de l'échantillon observé et est par conséquent intéressé par la distribution asymptotique non dégénérée de la moyenne d'échantillon en tant qu'estimateur. Nous avons vu dans la Section 4.3 que pour ce propos nous devrions examiner la distribution de  $n^{1/2}(m_1 - \mu)$ , où  $m_1$  est la moyenne d'échantillon. Typiquement, nous souhaitons étudier

$$n^{1/2}(m_1 - \mu) = n^{-1/2} \sum_{t=1}^n (y_t - \mu),$$

où  $y_t - \mu$  a une variance de  $\sigma_t^2$ .

Pour poursuivre, nous exposerons le **théorème de la limite centrale** simple suivant.

*Théorème 4.2. Théorème de la Limite Centrale Simple. (Lyapunov)*

Soit  $\{y_t\}$  une suite de variables aléatoires centrées avec des variances  $\sigma_t^2$  telles que  $\underline{\sigma}^2 \leq \sigma_t^2 \leq \bar{\sigma}^2$  pour deux constantes finies positives,  $\underline{\sigma}^2$  et  $\bar{\sigma}^2$ , et des moments d'ordre trois  $\mu_3$  tels que en valeur absolue  $\mu_3 \leq \bar{\mu}_3$  pour une constante finie  $\bar{\mu}_3$ . De plus

$$\sigma_0^2 \equiv \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \sigma_t^2 \right)$$

existe. Alors la suite

$$\left\{ n^{-1/2} \sum_{t=1}^n y_t \right\}$$

tend en distribution vers une limite caractérisée par une distribution normale d'espérance nulle et de variance  $\sigma_0^2$ .

Le Théorème 4.2 s'applique directement à l'exemple (4.26). Ainsi notre utilisateur hypothétique peut, dans les limites de la théorie asymptotique, utiliser la distribution  $N(0, \sigma_0^2)$  pour l'inférence statistique sur l'estimation

$m_1$  via la variable aléatoire  $n^{1/2}(m_1 - \mu)$ . La connaissance de  $\sigma_0^2$  n'est pas nécessaire, pourvu qu'on puisse l'estimer de manière convergente.

Bien que nous n'ayons pas l'intention d'apporter une preuve formelle à ce théorème de la limite centrale, étant donnés les points techniques qu'une telle preuve nécessiterait, il n'est pas difficile de donner une idée générale pour justifier la véracité du résultat. Pour faire simple, considérons le cas où toutes les variables  $y_t$  de la suite  $\{y_t\}$  ont la même distribution avec la variance  $\sigma^2$ . Ainsi, à l'évidence, la variable

$$S_n \equiv n^{-1/2} \sum_{t=1}^n y_t$$

est d'espérance nulle et de variance  $\sigma^2$  pour chaque  $n$ . Mais qu'en est-il des moments d'ordre supérieurs de  $S_n$ ? A titre d'exemple, considérons le moment d'ordre quatre. Il s'agit de

$$E(S_n^4) = \frac{1}{n^2} \sum_{r=1}^n \sum_{s=1}^n \sum_{t=1}^n \sum_{u=1}^n E(y_r y_s y_t y_u). \quad (4.27)$$

Puisque tous les  $y_t$  sont mutuellement indépendants et d'espérance nulle, chaque terme dans la somme quadruple de (4.27) est nul à moins que les indices soient tous identiques ou qu'il y ait des paires d'indices égaux (avec, par exemple,  $r = t$  et  $s = u$  où  $r \neq s$ ). Si tous les indices sont identiques, alors la valeur du terme correspondant est juste le quatrième moment de la distribution des  $y_t$ . Mais il n'y a que  $n$  termes de cette sorte. Avec le facteur de  $n^{-2}$  dans (4.27), nous voyons que la contribution de ces termes à (4.27) est de l'ordre de  $n^{-1}$ . Par ailleurs le nombre de termes pour lesquels les indices apparaissent par paires est  $3n(n-1)$ ,<sup>4</sup> qui est  $O(n^2)$ . Ainsi la contribution des derniers termes à (4.27) est de l'ordre de l'unité. Mais, et ceci est le point crucial de l'argumentation, la *valeur* de chacun de ces termes est juste le carré de la variance de chaque  $y_t$ , ou  $\sigma^4$ . Ainsi, à l'ordre dominant, le quatrième moment de  $S_n$  dépend seulement de la variance des  $y_t$ ; il *ne dépend pas* du quatrième moment de la distribution des  $y_t$ .<sup>5</sup>

Un argument similaire s'applique à tous les moments de  $S_n$  d'ordre supérieur à 2. Ainsi, à l'ordre dominant, tous ces moments ne dépendent que de la variance  $\sigma^2$  et d'aucune autre propriété de la distribution des  $y_t$ . Ceci étant, s'il est légitime de caractériser une distribution par ses moments, alors la distribution limite de la suite  $\{S_n\}_{n=1}^\infty$  ne dépend que de  $\sigma^2$ . Par

<sup>4</sup> Il existe trois manières de combiner les quatre indices,  $n$  manières de choisir l'indice de la première paire, et  $n-1$  manières de choisir un indice différent pour la seconde paire.

<sup>5</sup> La valeur de ce quatrième moment est  $n^{-2}$  fois  $3n(n-1)$  fois  $\sigma^4$ , dont l'ordre supérieur est  $3\sigma^4$ . Il s'agit du quatrième moment de la distribution normale.

conséquent, la distribution limite doit être *la même* pour toutes les distributions possibles où la variance de  $y_t$  est égale à  $\sigma^2$ , sans considération des autres propriétés de cette distribution. Ceci signifie que nous pouvons calculer la distribution limite en choisissant n'importe quelle distribution à condition qu'elle soit d'espérance 0 et de variance  $\sigma^2$ , le résultat étant indépendant de notre choix.

Le choix le plus simple est celui de la **distribution normale**,  $N(0, \sigma^2)$ . Le calcul de la distribution limite est très facile pour ce choix:  $S_n$  est juste une somme de  $n$  variables indépendantes normales, à savoir les  $n^{-1/2}y_t$ , toutes d'espérance 0 et de variance  $n^{-1}\sigma^2$ . Par conséquent,  $S_n$  est elle-même distribuée suivant une  $N(0, \sigma^2)$  pour tout  $n$ . Si la distribution est  $N(0, \sigma^2)$  pour tous les  $n$  indépendamment de  $n$ , alors la distribution limite est également  $N(0, \sigma^2)$ . Mais si c'est le cas pour une somme d'éléments normaux, nous pouvons conclure d'après notre argumentation précédente que la distribution limite de *n'importe quelle* suite  $S_n$  issue d'une somme d'éléments indépendants, d'espérance nulle et de variance  $\sigma^2$ , sera  $N(0, \sigma^2)$ .

La discussion qui précède a éludé de nombreux détails techniques, mais s'est concentrée sur l'essentiel qui fournit les preuves des théorèmes de la limite centrale. Nous pouvons assurer à nouveau que l'aspect le plus important du résultat de la limite centrale est que la distribution limite est *normale*.

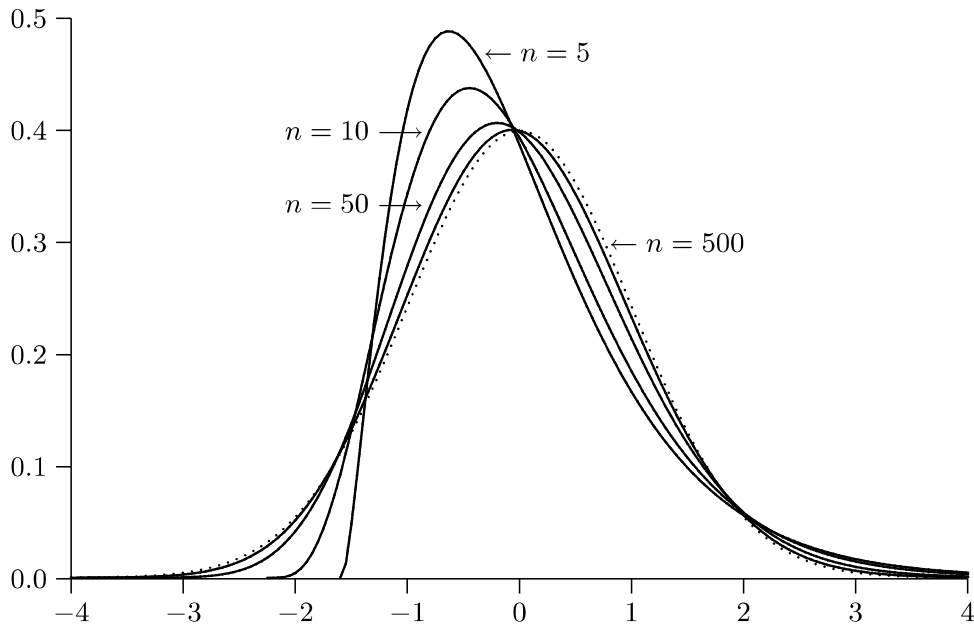
Dans la pratique, nous souhaitons typiquement estimer un vecteur de variables aléatoires, disons  $\beta$ . Notons  $\hat{\beta}^n$  l'estimation sur un échantillon de taille  $n$  et  $\beta_0$  la véritable valeur. Alors, comme nous le démontrerons dans le prochain chapitre, l'application d'un théorème de la limite centrale adéquat nous permettra généralement de conclure que  $n^{1/2}(\hat{\beta}^n - \beta_0)$  est asymptotiquement distribué suivant une normale multivariée avec un vecteur d'espérances nulles et une certaine matrice de covariance précise que l'on peut estimer de façon convergente.

Les théorèmes de la limite centrale sont utiles dans la pratique parce que, dans de nombreux cas, ils conduisent à de bonnes approximations même quand  $n$  n'est pas très grand. Ceci est illustré dans la Figure 4.1, qui traite délibérément d'un cas où les variables aléatoires sous-jacentes sont fortement non normales, de sorte qu'un théorème de la limite centrale s'y applique assez mal. Chacune des variables aléatoires sous-jacente  $y_t$  est distribuée suivant une  $\chi^2(1)$ , une distribution qui manifeste une asymétrie à droite très prononcée: le mode de la distribution est en zéro, aucune valeur n'est inférieure à zéro, et il y a une très longue queue du côté droit. La figure illustre la densité de

$$n^{-1/2} \sum_{t=1}^n \frac{y_t - \mu}{\sigma},$$

où, dans ce cas,  $\mu = 1$  et  $\sigma = \sqrt{2}$ , pour  $n = 5$ ,  $n = 10$ ,  $n = 50$ , et  $n = 500$ . Par comparaison, la densité d'une distribution normale centrée réduite a été tracée en pointillés. Il est clair que le théorème de la limite centrale fonctionne





**Figure 4.1** L'approximation normale pour des valeurs différentes de  $n$

très bien pour  $n = 500$  et assez bien pour  $n = 50$ , en dépit de la distribution fortement non normale des  $y_t$ . Dans de nombreux autres cas, par exemple quand les  $y_t$  sont distribués uniformément, la convergence vers la normalité asymptotique est beaucoup plus rapide.

Naturellement, tous les estimateurs ne sont pas convergents et asymptotiquement normaux. Parfois, aucun théorème de la limite centrale ne s'applique, et, plus rarement aucune loi des grands nombres ne s'applique non plus. En particulier, aucun théorème ni aucune loi ne s'applique quand la somme de variables aléatoires dans une suite aléatoire ne possède même pas un premier moment. Nous illustrons à présent un tel cas. Nous supposons que les variables aléatoires sont distribuées selon la **distribution de Cauchy**, caractérisée par sa fonction de densité,

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

La distribution de Cauchy est reliée à la distribution normale par le fait que le ratio de deux variables normales centrées réduites indépendantes est distribué suivant une loi de Cauchy; consulter l'Annexe B. La distribution n'a aucun moment d'ordre supérieur ou égal à 1, comme le montre l'intégrale

$$\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} dx \quad (4.28)$$

qui diverge en chacune de ses limites. Une primitive de la fonction à intégrer dans (4.28) est

$$\frac{1}{2\pi} \log(1+x^2),$$

qui tend vers l'infini quand  $x \rightarrow \pm\infty$ .

Une autre propriété de la distribution de Cauchy illustre mieux notre exemple actuel. Celle-ci indique que la moyenne d'un nombre fini quelconque de variables aléatoires de Cauchy indépendantes suit elle-même la distribution de Cauchy. Nous ne démontrerons pas cette propriété mais prendrons acte de certaines de ses conséquences.

Supposons que l'on dispose d'un échantillon aléatoire des observations qui sont des tirages d'une **distribution de Cauchy traduite**. C'est-à-dire qu'il existe un **paramètre de translation**  $\mu$ , qui ne correspond pas à l'espérance de la distribution, tel que pour chaque observation  $y_t$ , la quantité de  $y_t - \mu$  suit la distribution de Cauchy. Examinons à présent les propriétés de la moyenne d'échantillon  $m_1$ , qui naturellement existe toujours, comme estimateur du paramètre  $\mu$ . Nous avons

$$m_1 = \frac{1}{n} \sum_{t=1}^n y_t = \frac{1}{n} \sum_{t=1}^n (\mu + u_t),$$

où, par définition,  $u_t \equiv y_t - \mu$  est une variable aléatoire de Cauchy. Ainsi,

$$m_1 = \mu + \frac{1}{n} \sum_{t=1}^n u_t.$$

Le second terme du membre de droite de l'équation précédente est simplement une moyenne de variables de Cauchy indépendantes, et par conséquent, elle suit la distribution de Cauchy conformément au résultat précité. A l'évidence,  $m_1$  n'est pas une estimation convergente de  $\mu$ , puisque pour *n'importe quel*  $n$ ,  $m_1$  a une distribution de Cauchy traduite. C'est comme si le facteur  $1/n$  était le facteur approprié, non pas pour la loi des grands nombres, mais pour une forme du théorème de la limite centrale, puisqu'il s'agit de la somme des  $u_t$  divisée par  $n$  plutôt que par  $n^{1/2}$  qui possède une distribution asymptotique non dégénérée. Bien sûr, cette distribution asymptotique est une distribution de Cauchy plutôt qu'une distribution normale. Bien que  $m_1$  ne soit pas un estimateur satisfaisant du paramètre  $\mu$ , puisque  $\mu$  est à la fois la médiane et le mode de la distribution, d'autres manières de l'estimer existent, et de fait la médiane d'échantillon, par exemple, en est une.

## 4.7 QUELQUES RÉSULTATS UTILES

Cette section est destinée à servir de référence pour la majeure partie du reste de ce livre. Nous nous contenterons de dresser une liste (avec un commentaire éventuel mais sans démonstration) de définitions et de théorèmes utiles. A la fin de celle-ci nous présenterons deux ensembles de conditions de régularité qui posséderont chacune un ensemble d'implications intéressant. Par la suite,

nous serons capables de poser des hypothèses selon lesquelles l'un ou l'autre de ces deux ensembles de conditions de régularité est satisfait et par la même de dresser une liste importante de conditions utiles sans plus de cérémonie.

Pour commencer, nous nous concentrerons sur les lois des grands nombres et sur les propriétés qui leurs permettent d'être satisfaites. Dans tous ces théorèmes, nous considérons une suite de sommes  $\{S_n\}$  où

$$S_n \equiv \frac{1}{n} \sum_{t=1}^n y_t.$$

Nous nous référerons aux variables aléatoires  $y_t$  sous le nom de **termes** (aléatoires). Tout d'abord, nous présentons un théorème possédant très peu de contraintes portant sur les moments des termes aléatoires mais des contraintes sur leur homogénéité.

*Théorème 4.3. (Khinchin)*

Si les variables aléatoires  $y_t$  de la suite  $\{y_t\}$  sont mutuellement indépendantes et toutes distribuées suivant la même distribution, d'espérance  $\mu$ , alors

$$\Pr\left(\lim_{n \rightarrow \infty} S_n = \mu\right) = 1.$$

Seule l'existence du moment d'ordre un est nécessaire, mais tous les termes doivent être identiquement distribués. Notons que la constance de l'espérance des termes signifie qu'il n'est pas nécessaire de centrer les variables  $y_t$ .

Ensuite, nous présentons un théorème dû à Kolmogorov, qui nécessite à nouveau l'indépendance des termes, et à présent leurs moments d'ordre deux, mais très peu de contraintes supplémentaires du point de vue de leur homogénéité.

*Théorème 4.4. (Kolmogorov)*

Soit la suite  $\{y_t\}$  de variables aléatoires centrées mutuellement indépendantes possédant la propriété

$$\lim_{n \rightarrow \infty} \left( n^{-2} \sum_{t=1}^n \text{Var}(y_t) \right) < \infty.$$

Alors  $S_n \rightarrow 0$  presque sûrement.

Ceci constitue un résultat très fort, puisqu'à cause du facteur  $n^{-2}$  il n'est pas difficile de satisfaire la condition spécifiée. Des variances bornées la satisfont aisément, par exemple.

Dans les théorèmes suivants, l'hypothèse de l'indépendance des termes est relâchée. Une certaine régularité est naturellement toujours nécessaire, et pour cela nous avons besoin de quelques définitions. Rappelons à ce point

la définition d'une fonction indicatrice, qui sera utilisée pour un vecteur de variables aléatoires: si la variable aléatoire  $\mathbf{y}$  est observée dans  $\mathbb{R}^k$  et  $G$  est un sous-ensemble quelconque de  $\mathbb{R}^k$  sur lequel  $\Pr(\mathbf{y} \in G)$  est bien définie, alors

$$I_G(\mathbf{y}) = \begin{cases} 1 & \text{si } \mathbf{y} \in G \\ 0 & \text{sinon.} \end{cases}$$

Ensuite, nous définissons la notion importante d'**espérance conditionnelle**. Nous utiliserons considérablement ce concept tout au long de cet ouvrage.

*Définition 4.9.*

L'espérance de la variable aléatoire  $y$  conditionnellement au vecteur de variables aléatoires  $\mathbf{z}$  est une variable aléatoire  $w$  qui est une fonction déterministe des variables conditionnantes  $\mathbf{z}$  et qui possède la propriété suivante de définition. Pour tout  $G \subseteq \mathbb{R}^k$  tel que  $\Pr(\mathbf{z} \in G)$  est bien définie,

$$E(wI_G(\mathbf{z})) = E(yI_G(\mathbf{z})). \quad (4.29)$$

L'espérance conditionnelle  $w$  est notée  $E(y | \mathbf{z})$ .

Observons qu'une espérance conditionnelle est une *variable aléatoire*, en tant que fonction des variables conditionnantes  $\mathbf{z}$ . L'espérance ordinaire (non conditionnelle), qui naturellement n'est pas aléatoire, peut être considérée comme l'espérance conditionnelle à une variable non stochastique. Par ailleurs, l'espérance d'une variable conditionnellement à elle-même est la variable elle-même.

Une espérance calculée conditionnellement à un ensemble de variables conditionnantes  $\mathbf{z}$  sera la même que l'espérance calculée conditionnellement à un autre ensemble  $\mathbf{z}'$  s'il existe une bijection partant de l'ensemble  $\mathbf{z}$  dans l'ensemble  $\mathbf{z}'$ , de sorte que toute fonction de  $\mathbf{z}$  puisse se transformer en une fonction de  $\mathbf{z}'$ . Une conséquence immédiate est que l'espérance d'une fonction  $h(y)$  d'une variable aléatoire  $y$ , conditionnellement à  $y$ , est juste  $h(y)$ .

Une autre conséquence importante de la définition d'une espérance conditionnelle est ce que l'on appelle la **loi des espérances itérées**, que l'on peut exposer comme suit:

$$E(E(y | \mathbf{z})) = E(y).$$

La démonstration de cette loi est une conséquence immédiate de l'utilisation de l'ensemble  $\mathbb{R}^k$  pour  $G$  dans (4.29).

Les définitions suivantes sont plutôt techniques, comme le sont les exposés des lois des grands nombres qui les utilisent. Certains lecteurs peuvent par conséquent souhaiter passer directement aux définitions des deux ensembles de conditions de régularité, que nous appelons WULLN et CLT, présentées à la fin de cette section. Ces lecteurs peuvent revenir à ces définitions techniques lorsque nous y ferons référence dans la suite du livre.

*Définition 4.10.*

La suite  $\{y_t\}$  est dite **stationnaire** si pour tout  $k$  fini la distribution jointe de l'ensemble lié  $\{y_t, y_{t+1}, \dots, y_{t+k}\}$  est indépendante de  $t$ .

*Définition 4.11.*

La suite stationnaire  $\{y_t\}$  est dite **ergodique** si, pour deux applications bornées quelconques  $Y: \mathbb{R}^k \rightarrow \mathbb{R}$  et  $Z: \mathbb{R}^l \rightarrow \mathbb{R}$ ,

$$\begin{aligned} & \lim_{n \rightarrow \infty} |E(Y(y_i, \dots, y_{i+k})Z(y_{i+n}, \dots, y_{i+n+l}))| \\ &= |E(Y(y_i, \dots, y_{i+k}))| |E(Z(y_i, \dots, y_{i+l}))|. \end{aligned}$$

*Définition 4.12.*

La suite  $\{y_t\}$  est dite **uniformément mélangeante**, ou  **$\phi$ -mélangeante**, s'il existe une suite de nombres positifs  $\{\phi_n\}$ , convergeant vers zéro, tels que, pour deux applications bornées quelconques  $Y: \mathbb{R}^k \rightarrow \mathbb{R}$  et  $Z: \mathbb{R}^l \rightarrow \mathbb{R}$ ,

$$|E(Y(y_t, \dots, y_{t+k}) | Z(y_{t+n}, \dots, y_{t+n+l})) - E(Y(y_t, \dots, y_{t+k}))| < \phi_n.$$

Le symbole  $E(\cdot | \cdot)$  désigne une espérance conditionnelle, comme précédemment définie.

*Définition 4.13.*

La suite  $\{y_t\}$  est dite  **$\alpha$ -mélangeante** s'il existe une suite de nombres positifs  $\{\alpha_n\}$ , convergeant vers zéro, tels que, si  $Y$  et  $Z$  sont des applications comparables à celle de la définition précédente, alors

$$|E(Y(y_t, \dots, y_{t+k})Z(y_{t+n}, \dots, y_{t+n+l})) - E(Y(\cdot))E(Z(\cdot))| < \alpha_n.$$

Les trois dernières définitions peuvent s'interpréter comme des définitions de formes variées d'**indépendance asymptotique**. Selon elles, les variables aléatoires  $y_t$  et  $y_s$  sont d'autant plus indépendantes (dans un certain sens) que les indices  $t$  et  $s$  sont éloignés. Pour en savoir plus sur ces concepts et sur leurs implications, consulter White (1984) et Spanos (1986, Chapitre 8). Un aspect utile des propriétés de mélange et d'ergodicité est que si la suite  $\{y_t\}$  possède une de ces propriétés, c'est également le cas des suites de fonctions des  $y_t$ ,  $\{Y(y_t)\}$ , de la même manière que les fonctions de deux variables aléatoires indépendantes sont elles-mêmes indépendantes. Un résultat plus fort est aussi vrai. Une suite de la forme  $\{Y(y_t, \dots, y_{t+i})\}$  préserve aussi les propriétés de mélange et d'ergodicité et, dans le cas de la propriété d'ergodicité, l'étendue de la dépendance  $i$  peut être infinie. La propriété  $\phi$ -mélangeante est la plus forte: elle implique la propriété  $\alpha$ -mélangeante, qui implique à son tour l'ergodicité pour les suites stationnaires.

Les propriétés de mélange sont importantes si l'on désire gérer des suites non stationnaires. Dans ce livre, nous aborderons peu le sujet (sauf dans le Chapitre 20), et préférons par conséquent présenter le prochain théorème consacré aux suites *stationnaires* et possédant la condition de régularité la plus faible, à savoir l'ergodicité. Ce théorème, dû au célèbre mathématicien américain G. D. Birkhoff, est de fait bien connu dans la littérature mathématique et dans de nombreuses disciplines autres que l'économétrie.

*Théorème 4.5. Théorème Ergodique.*

Si la suite stationnaire  $\{y_t\}$  est ergodique dans le sens de la Définition 4.11 et si l'espérance  $\mu$  de  $y_t$  existe et est finie, alors  $S_n$  tend vers  $\mu$  presque sûrement quand  $n \rightarrow \infty$ .

A nouveau dans ce théorème aucun centrage n'est nécessaire, puisque la propriété de stationnarité assure que tous les  $y_t$  ont la *même* espérance.

Certaines définitions supplémentaires sont nécessaires avant d'exposer le prochain théorème.

*Définition 4.14.*

Une suite  $\{y_t\}$  de variables aléatoires est appelée **martingale** si, pour tout  $t$ ,  $E(|y_t|)$  existe et est finie et si, pour tout  $t$ ,

$$E(y_{t+1} | y_t, \dots, y_1) = y_t.$$

Les martingales sont des types de suites très importants. Un exemple simple est celui d'une suite  $\{Z_n\}$  des sommes des variables aléatoires centrées indépendantes  $y_t$ :

$$E(Z_{n+1} | Z_n, \dots, Z_1) = E\left(\sum_{t=1}^{n+1} y_t | Z_n, \dots, Z_1\right) = E\left(\sum_{t=1}^{n+1} y_t | y_n, \dots, y_1\right),$$

puisque chacun des ensembles  $\{Z_n, \dots, Z_1\}$  et  $\{y_n, \dots, y_1\}$  détermine l'autre de manière unique. Alors, comme requis,

$$\begin{aligned} E\left(\sum_{t=1}^{n+1} y_t | y_n, \dots, y_1\right) &= E(y_{n+1} | y_n, \dots, y_1) + \sum_{t=1}^n y_t \\ &= E(y_{n+1}) + \sum_{t=1}^n y_t = Z_n. \end{aligned}$$

Les martingales apparaissent telles quelles de temps à autre en économétrie, mais une notion plus applicable est d'une suite de différences de martingale.

*Définition 4.15.*

Une suite  $\{y_t\}$  est dite **suite de différences de martingale** si

$$E(y_{t+1} | y_t, \dots, y_1) = 0.$$

Cette définition est très courte parce que la condition implique non seulement l'existence des espérances non conditionnelles  $E(y_t)$  mais également que celles-ci soient nulles, la suite étant alors centrée. Consulter Spanos (1986, Chapitre 8).

*Théorème 4.6. (Chow)*

Si  $\{y_t\}$  est une suite de différences de martingale et s'il existe un  $r \geq 1$  tel que la série

$$\sum_{t=1}^{\infty} t^{-(1+r)} E(|y_t|^{2r})$$

converge, alors  $S_n \rightarrow 0$  presque sûrement.

La condition de régularité est très faible du fait du facteur  $t^{-(1+r)}$ . Notons que

$$\sum_{t=1}^{\infty} t^{-(1+r)}$$

converge pour tout  $r > 0$ . En particulier la condition est satisfaite si les valeurs absolues des moments d'ordre  $(2k)$  des  $y_t$ ,  $E(|y_t|^{2r})$ , sont uniformément bornées, ce qui signifie qu'il existe une constante  $K$ , indépendante de  $t$ , telle que  $E(|y_t|^{2r}) < K$  pour tout  $t$ . Consulter Stout (1974) et Y. S. Chow (1960, 1967).

Nous sommes maintenant prêts pour aborder une sélection de théorèmes de la limite centrale. Une procédure utile, comparable au centrage utilisé si souvent dans notre discussion sur les lois des grands nombres, est le **centrage et la réduction** d'une suite. Pour que ceci soit possible, chaque variable de la suite  $\{y_t\}$  doit avoir à la fois un premier et un second moment. Alors si  $\mu_t$  et  $v_t$  désignent, respectivement, l'espérance et la variance de  $y_t$ , la suite d'élément type  $z_t \equiv (y_t - \mu_t)/\sqrt{v_t}$  est dite centrée réduite. Ainsi, chaque variable d'une telle suite a une espérance nulle et une variance unitaire. Pour ce qui concerne notre collection de théorèmes de la limite centrale, la variable  $S_n$  associée à une suite  $\{y_t\}$  sera redéfinie comme suit, où  $\mu_t$  et  $v_t$  sont, respectivement, l'espérance et la variance de  $y_t$ :

$$S_n \equiv \frac{\sum_{t=1}^n (y_t - \mu_t)}{(\sum_{t=1}^n v_t)^{1/2}}.$$

Il est clair que  $\{S_n\}$  est centrée réduite si les  $y_t$  sont indépendantes.

*Théorème 4.7. (Lindeberg-Lévy)*

Si les variables de la suite aléatoire  $\{y_t\}$  sont indépendantes et de même distribution avec une espérance  $\mu$  et une variance  $v$ , alors  $S_n$  converge en distribution vers la distribution normale centrée réduite  $N(0, 1)$ .

Ce théorème repose sur des contraintes minimales pour les moments des variables mais sur des contraintes plus fortes sur leur homogénéité. Notons que, dans ce cas,

$$S_n = (nv)^{-1/2} \sum_{t=1}^n y_t.$$

Le prochain théorème autorise une hétérogénéité plus forte mais nécessite toujours l'indépendance.

*Théorème 4.8. (Lyapunov)*

Pour chaque entier positif  $n$ , imaginons la suite finie  $\{y_t^n\}_{t=1}^n$  composée de variables aléatoires centrées indépendantes possédant les variances  $v_t^n$ . Soit  $s_n^2 \equiv \sum_{t=1}^n v_t^n$  et supposons la **condition de Lindeberg** satisfaite, à savoir que pour tout  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \left( \sum_{t=1}^n s_n^{-2} E((y_t^n)^2 I_G(y_t^n)) \right) = 0,$$

où l'ensemble  $G$  utilisé comme fonction indicatrice est  $\{y : |y| \geq \varepsilon s_n\}$ . Alors  $s_n^{-1} \sum_{t=1}^n y_t^n$  converge en distribution vers une  $N(0, 1)$ .

Notre dernier théorème de la limite centrale autorise des suites dépendantes.

*Théorème 4.9. (McLeish)*

Pour tout entier positif  $n$ , soit les suites finies  $\{y_t^n\}_{t=1}^n$  des suites de différences de martingale avec  $v_t^n \equiv \text{Var}(y_t^n) < \infty$ , et  $s_n^2 \equiv \sum_{t=1}^n v_t^n$ . Si pour tout  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \left( s_n^{-2} \sum_{t=1}^n E((y_t^n)^2 I_G(y_t^n)) \right) = 0,$$

où  $G \equiv \{y : |y| \geq \varepsilon s_n\}$  une fois de plus, et si la suite

$$\left\{ \sum_{t=1}^n \frac{(y_t^n)^2}{n^{-1} s_n^2} \right\}$$

obéit à une loi des grands nombres et converge par conséquent vers 1, alors  $s_n^{-1} \sum_{t=1}^n y_t^n$  converge en distribution vers  $N(0, 1)$ .

Consulter McLeish (1974). Observons la condition supplémentaire requise dans ce théorème, qui garantit que la variance de la distribution limite est la même que la limite des variances des variables dans  $s_n^{-1} \sum_{t=1}^n y_t^n$ .

Nous pouvons à présent exposer notre collection de conditions de régularité adéquates pour les besoins de ce chapitre et des suivants. Il est commode de commencer par les conditions CLT, dont le nom vient du fait qu'elles comprennent un théorème de la limite centrale.



*Définition 4.16.*

Une suite  $\{y_t\}$  de variables aléatoires centrées satisfait la **condition CLT** si elle satisfait un théorème de la limite centrale tel que: soit  $\text{Var}(y_t) = \sigma_t^2$  et  $s_n^2 \equiv \sum_{t=1}^n \sigma_t^2$ , avec

$$\text{plim}_{n \rightarrow \infty} \left( s_n^{-2} \sum_{t=1}^n y_t^2 \right) = 1.$$

Alors  $s_n^{-1} \sum_{t=1}^n y_t$  tend en distribution vers  $N(0, 1)$ .

Nous ne spécifions pas le théorème de la limite centrale qui justifie la conclusion; nous demandons seulement qu'un théorème de la sorte la justifie. La condition supplémentaire est imposée de sorte qu'il est possible d'obtenir des estimations efficaces des variances des variables aléatoires asymptotiquement normales. Plus précisément, elle nécessite que  $n^{-1} \sum_{t=1}^n y_t^2$  soit une estimation convergente de la variance de  $n^{-1/2} \sum_{t=1}^n y_t$ .

Dans la pratique, nous serons souvent amenés à appliquer la condition CLT à une **suite vectorielle**. Par exemple, si nous avons une fonction de plusieurs paramètres qui retourne une valeur scalaire, il peut être intéressant d'appliquer la condition CLT au vecteur de ses dérivées partielles. Dans ce contexte, le théorème suivant s'avère extrêmement utile.

*Théorème 4.10. (Normalité Multivariée)*

Si une collection  $\{z_1, \dots, z_m\}$  de variables aléatoires distribuées normalement a la propriété que n'importe quelle combinaison linéaire de l'ensemble est aussi distribuée normalement, alors les variables dans  $\{z_1, \dots, z_m\}$  ont une distribution jointe normale multivariée.

Consulter Rao (1973), parmi d'autres, pour des références et une démonstration. Les arguments qui conduisent à une conclusion de normalité asymptotique pour une collection finie de variables aléatoires prises séparément s'appliquera presque toujours aux combinaisons linéaires des éléments de somme, de sorte que la notion de **normalité multivariée** d'une collection finie de variables est habituellement insuffisante. Nous ne prendrons parfois même pas la peine de mentionner le problème dans la suite du livre et ferons référence à la condition CLT s'appliquant directement à une suite vectorielle.

La seconde de nos deux collections de conditions introduit une idée nouvelle, celle de **convergence uniforme** d'une suite, que l'on peut employer si les éléments de la suite sont des *fonctions* des variables ou paramètres (non aléatoires). Cette situation se manifestera régulièrement lorsque nous examinerons dans le prochain chapitre les procédures d'estimation. Nous traiterons des suites de variables aléatoires qui dépendent de paramètres de modèle inconnus et qui devront par conséquent satisfaire une loi des grands nombres pour n'importe quel ensemble de valeurs de ces paramètres dans un voisinage quelconque. La convergence uniforme est un renforcement de la notion de convergence qui permet de tirer des conclusions telle la continuité ou

l'intégrabilité par rapport aux paramètres des fonctions limites si les fonctions dans la suite sont elles-mêmes continues ou intégrables.

Nous donnerons la définition formelle plus tard dans le but d'introduire des références ultérieures. Les détails de la définition ne sont pas importants pour l'instant: la clé du problème est qu'un *certain* renforcement de la propriété de convergence est nécessaire si les fonctions en tant que limites des suites de fonctions héritent des propriétés utiles de continuité et d'intégrabilité que possèdent les éléments de ces suites.

*Définition 4.17.*

Une suite de fonctions aléatoires  $\{y_t(\beta)\}$  d'un vecteur des arguments  $\beta \in \mathbb{R}^k$  satisfait la **condition WULLN** (loi faible uniforme des grands nombres) dans un voisinage quelconque  $R$  contenu dans  $\mathbb{R}^k$  si les espérances  $E(y_t(\beta))$  existent pour tout  $t$  et  $\beta \in R$ , si

$$\bar{y}(\beta) \equiv \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n E(y_t(\beta)) \right) \quad (4.30)$$

existe et est finie pour tout  $\beta \in R$ , et si la convergence dans (4.30) est uniforme dans le sens suivant: pour tout  $\varepsilon > 0$  il existe un  $N$  tel que, pour tout  $\beta \in R$ ,

$$\left| \frac{1}{n} \sum_{t=1}^n E(y_t(\beta)) - \bar{y}(\beta) \right| < \varepsilon \quad \text{pour tout } n > N, \text{ et} \quad (4.31)$$

$$\Pr \left( \max_{\beta \in R} \left| \frac{1}{n} \sum_{t=1}^n y_t(\beta) - \bar{y}(\beta) \right| > \varepsilon \right) < \varepsilon \quad \text{pour tout } n > N.$$

Nous tenons compte explicitement dans cette condition de la possibilité que les distributions des  $y_t$  utilisées pour calculer les espérances dans (4.31) et (4.30) dépendent elles-mêmes de  $\beta$ .

## 4.8 CONCLUSION

Nous avons essayé dans ce chapitre de fournir une approche intuitive des outils, mathématiques et probabilistes, employés dans la théorie asymptotique. La plupart des matières, et en particuliers celles de la Section 4.7, ne nécessitent pas une maîtrise parfaite pour l'instant. Elles sont là au cas où nous en aurions besoin plus tard dans le livre, quand leur objet sera plus évident. Les concepts absolument essentiels de ce chapitre sont ceux des lois des grands nombres et des théorèmes de la limite centrale. Il *est* nécessaire de maîtriser intuitivement ces derniers avant d'aborder le prochain chapitre en toute sécurité. Les idées de convergence et de normalité asymptotique sont

probablement plus familières, et nous les détaillerons de toute manière dans le prochain chapitre.

Notre sélection des théorèmes présentés dans ce chapitre s'inspire largement des ouvrages de Billingsley (1979) et White (1984). Consulter aussi Stout (1974) et Lukacs (1975). Cependant, ceux-ci ne sont pas tout à fait des textes élémentaires et sont recommandés pour des lectures approfondies plutôt que pour la clarification d'un quelconque concept qui serait, malgré tous nos efforts, toujours obscur. Spanos (1986) fournit un traitement de la plupart de ces matières à un niveau technique plus abordable.

## TERMES ET CONCEPTS

atomes, pour une c.d.f.	mélangeante uniforme
condition CLT	moments (d'une distribution)
condition de Lindeberg	moments d'échantillon
condition WULLN	normalité asymptotique
convergence en distribution (ou en loi)	normalité multivariée
convergence en probabilité	ordonnée (suite)
convergence, forte ou faible	paramètre de translation
convergence presque sûre	passage à la limite
convergence uniforme (d'une suite)	population
distribution de Cauchy, ordinaire et translatée	processus stochastique
distribution dégénérée	quand $n$ tend vers l'infini
distribution normale	règle (pour définir un processus stochastique ou DGP)
distribution sans biais (et convergence)	relation de même ordre
distribution symétrique	relation d'ordre inférieur
données chronologiques (temporelles)	relations d'ordre stochastique
données de panel	suite centrée
données en coupe transversale	suite centrée réduite (procédure)
échantillon aléatoire	suite de différence de martingale
égalité asymptotique	suite réelle
ergodique (suite)	suite stationnaire
espace topologique	suite vectorielle
espérance conditionnelle	suites, convergentes et divergentes
fonction de distribution empirique	successeur (dans une suite)
fonction indicatrice	symboles $O, o$
indépendance asymptotique	symboles d'ordre, ordinaires et stochastiques
inégalité de Chebyshev	taille d'échantillon
limite en probabilité (plim)	taux de convergence
limite en probabilité non dégénérée	tendance temporelle
limite en probabilité non stochastique	théorème de la limite centrale
limite presque sûre	théorème d'existence de Kolmogorov
limites, finies et infinies	Théorème Fondamental de la Statistique
loi (d'une variable aléatoire)	théorie asymptotique
loi des espérances itérées	tirage aléatoire
loi des grands nombres, faible et forte	topologie naturelle
martingale	variable aléatoire
mélangeante (propriété d'une suite)	

# Chapitre 5

## Méthodes Asymptotiques et Moindres Carrés non Linéaires

### 5.1 INTRODUCTION

Dans le chapitre précédent, nous avons introduit la plupart des idées fondamentales de l'analyse asymptotique et établi certains résultats essentiels à partir de la théorie des probabilités. Dans ce chapitre, nous utilisons ces concepts et résultats pour démontrer un certain nombre de propriétés importantes de l'estimateur des moindres carrés non linéaires.

Dans la prochaine section, nous discutons du concept de l'**identification asymptotique** des **modèles paramétrisés** et, en particulier, des modèles estimés par NLS. Dans la Section 5.3, nous nous focaliserons sur la **convergence** de l'estimateur NLS pour des modèles identifiés asymptotiquement. Dans la Section 5.4, nous abordons sa **normalité asymptotique** et nous dérivons également la matrice de covariance asymptotique de l'estimateur NLS. Ceci nous conduit, dans la Section 5.5, à l'**efficacité asymptotique** des NLS, que nous démontrons par une extension au cas non linéaire du célèbre Théorème de Gauss-Markov pour les modèles de régression linéaire. Dans la Section 5.6, nous traitons de différentes propriétés utiles des résidus NLS. Enfin, dans la Section 5.7, nous considérons les distributions asymptotiques des statistiques de test introduites dans la Section 3.6 pour tester des restrictions sur des paramètres du modèle.

### 5.2 IDENTIFICATION ASYMPTOTIQUE

Quand nous parlons en économétrie de modèles qu'il s'agit d'estimer ou de tester, nous faisons référence à des ensembles de DGP. Quand nous abordons la théorie asymptotique, les DGP en question doivent être des processus stochastiques, pour les raisons exposées dans le Chapitre 4. Sans plus de cérémonie, notons  $\mathbb{M}$  un modèle qu'il s'agit d'estimer, de tester, ou les deux, et  $\mu$  un DGP type appartenant à  $\mathbb{M}$ . Ce que nous signifions très précisément par cette notation devra transparaître par la suite.

Le modèle le plus simple en économétrie est le modèle de régression linéaire, pour lequel il peut exister de nombreuses spécifications. L'une d'elles consiste à écrire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (5.01)$$

où  $\mathbf{y}$  et  $\mathbf{u}$  sont des vecteurs de dimension  $n$  et  $\mathbf{X}$  est une matrice déterministe de dimension  $n \times k$ . Ensuite des hypothèses (peut-être implicites) sont formulées sur la possibilité de définir  $\mathbf{X}$  par une règle quelconque (consulter la Section 4.2) pour tous les entiers positifs  $n$  supérieurs à une certaine valeur appropriée, de manière que, pour de telles valeurs de  $n$ ,  $\mathbf{y}$  suive la distribution  $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ . Cette distribution est unique si les paramètres  $\boldsymbol{\beta}$  et  $\sigma^2$  sont spécifiés. Nous pouvons par conséquent dire que le DGP est **complètement caractérisé** par les paramètres du modèle. Autrement dit, la connaissance des paramètres du modèle  $\boldsymbol{\beta}$  et  $\sigma^2$  identifie de façon unique un élément  $\mu$  de  $\mathbb{M}$ .

Par ailleurs, le modèle de régression linéaire peut aussi s'écrire comme

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (5.02)$$

sans aucune hypothèse de normalité. De nombreux aspects de la théorie des régressions linéaires sont applicables autant à (5.02) qu'à (5.01); par exemple, l'estimateur OLS est sans biais, et sa matrice de covariance est  $\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ . Mais la distribution du vecteur  $\mathbf{u}$ , et par conséquent aussi celle de  $\mathbf{y}$ , est maintenant **caractérisée partiellement** même si  $\boldsymbol{\beta}$  et  $\sigma^2$  sont connus. Par exemple, les aléas  $u_t$  pourraient être biaisés vers la gauche ou vers la droite, pourraient posséder des quatrièmes moments supérieurs ou inférieurs à  $3\sigma^4$ , ou pourraient même ne posséder aucun moment d'ordre supérieur à, par exemple, 6. Les DGP qui comportent toutes sortes de propriétés, dont certaines sont très étranges, constituent des cas particuliers du modèle de régression linéaire si celui-ci est défini par (5.02) plutôt que par (5.01).

Nous pouvons noter  $\mathbb{M}_1$  et  $\mathbb{M}_2$  les ensembles de DGP associés à (5.01) et à (5.02), respectivement. Ces ensembles de DGP sont différents,  $\mathbb{M}_1$  étant de fait un sous-ensemble propre de  $\mathbb{M}_2$ . Bien que pour tout DGP  $\mu \in \mathbb{M}_2$  il existe un  $\boldsymbol{\beta}$  et un  $\sigma^2$  qui correspondent à, et qui caractérisent partiellement,  $\mu$ , la relation inverse n'existe pas. Pour un  $\boldsymbol{\beta}$  et un  $\sigma^2$  donnés, il existe une infinité de DGP dans  $\mathbb{M}_2$  (dont un seul appartient à  $\mathbb{M}_1$ ) qui correspondent tous aux mêmes  $\boldsymbol{\beta}$  et  $\sigma^2$ . Ainsi, nous devons pour nos propos immédiats considérer (5.01) et (5.02) comme des modèles différents bien que les paramètres  $\mathbf{y}$  soient identiques.

La grande majorité des procédures statistiques et économétriques pour l'estimation des modèles utilise, comme c'est le cas pour le modèle de régression linéaire, les **paramètres du modèle**. Typiquement, c'est l'estimation de ces paramètres qui nous intéresse. Comme pour le modèle de régression linéaire, les paramètres peuvent caractériser totalement ou partiellement un DGP du modèle. Dans les deux cas, il doit être possible d'associer un vecteur paramétrique de façon unique à tout DGP  $\mu$  appartenant au modèle  $\mathbb{M}$ , même si le vecteur paramétrique en question est associé à de nombreux DGP.

La prise en compte dans notre notation de l'association entre les DGP d'un modèle et les paramètres du modèle sera pratique. En conséquence, nous définissons l'**application définissante des paramètres**  $\theta$  du modèle  $\mathbb{M}$ . Ainsi,  $\theta(\mu)$  désignera le vecteur paramétrique associé au DGP  $\mu$ . Par exemple, si  $\mathbb{M}$  est un modèle de régression linéaire, ou de tout autre type, et  $\mu$  un DGP appartenant au modèle, alors  $\theta(\mu) = (\beta, \sigma^2)$  pour les valeurs appropriées des paramètres  $\beta$  de la fonction de régression et de la variance des aléas  $\sigma^2$ . Le lecteur peut s'interroger sur la pertinence du terme compliqué "application définissante des paramètres" par rapport au terme plus simple "paramétrisation." La raison est que, en théorie formelle mathématique, une paramétrisation est une application qui opère dans l'autre sens; par conséquent, dans ce cas, elle associerait un DGP à un vecteur de paramètres donné. Puisque nous souhaitons typiquement donner la possibilité à un seul vecteur de paramètres de faire référence à un *ensemble* de DGP, nous avons préféré le terme plus compliqué.

En général, l'application  $\theta$  opère du modèle  $\mathbb{M}$  vers un **espace paramétrique**  $\Theta$ , qui sera habituellement soit  $\mathbb{R}^k$  soit un sous-ensemble de  $\mathbb{R}^k$ . Ici  $k$  est un entier positif: il indique la **dimension** de l'espace paramétrique  $\Theta$ . La relation qui lie l'application  $\theta$ , son **espace de départ**  $\mathbb{M}$ , et son **espace d'arrivée**  $\Theta$  est notée  $\theta: \mathbb{M} \rightarrow \Theta$ . Nous pouvons écrire  $\theta_0 \equiv \theta(\mu_0)$  si le vecteur de paramètres associé au DGP  $\mu_0$  est  $\theta_0$ . Si nous faisons référence à un DGP particulier, tel que  $\mu_0$  ou  $\mu_1$ , nous pouvons adopter notre pratique habituelle et écrire simplement  $\theta_0$  à la place de  $\theta(\mu_0)$  ou  $\theta_1$  à la place de  $\theta(\mu_1)$ . Nous utiliserons la notation  $(\mathbb{M}, \theta)$  pour un modèle et pour l'application définissante des paramètres qui lui est associée, et appellerons la paire  $(\mathbb{M}, \theta)$  **modèle paramétrisé**.

L'introduction de l'application  $\theta$  nous permet de traiter, dans le contexte asymptotique, la question de l'**identification** du modèle paramétrisé  $(\mathbb{M}, \theta)$  ou, plus précisément ici, des paramètres du modèle  $\theta$ . Avant de présenter la définition formelle de l'identification asymptotique, livrons-nous à une remarque préliminaire. Le simple fait que  $\theta$  soit définie comme une application de  $\mathbb{M}$  vers  $\Theta$  signifie qu'un seul vecteur de paramètres peut être associé à un DGP  $\mu$  donné. Ainsi, nous avons écarté à la source toute possibilité de gérer des modèles de régression avec une fonction de régression comme (2.07). Un exemple d'un tel modèle est

$$y_t = \beta_1 + \beta_2 X_{t2}^{\beta_3} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \quad (5.03)$$

Dans ce cas, si  $\beta_2 = 0$ , un seul et même DGP est associé à un ensemble entier de vecteurs de paramètres, puisque le choix de la valeur de  $\beta_3$  est alors sans conséquence. De façon similaire, si  $\beta_3 = 0$ , il n'existera aucun moyen d'identifier  $\beta_1$  et  $\beta_2$  séparément, et à nouveau un nombre infini de vecteurs paramétriques peut être associé au même DGP. Puisque des modèles comme (5.03) surviennent assez fréquemment dans les travaux appliqués, il est important de se souvenir que les résultats que nous dériverons dans ce chapitre

ne s'y appliquent pas, tout au moins sans l'imposition de certaines conditions supplémentaires. Dans le cas de (5.03), une solution simple consiste à définir l'espace paramétrique  $\Theta$  du modèle de manière à exclure les valeurs  $\beta_2 = 0$  et  $\beta_3 = 0$ , de sorte qu'il existe une application définissante des paramètres ayant de bonnes propriétés pour le modèle restreint de cette manière. En d'autres occasions, il sera peut-être possible de trouver une reparamétrisation du modèle pour laquelle une application définissante des paramètres existe.

Dans la Section 2.3, pour des modèles qu'il fallait estimer par moindres carrés non linéaires, un modèle était dit identifié par un certain ensemble de données si la fonction somme-des-carrés du modèle avait un unique minimum global atteint avec un unique vecteur paramétrique, compte tenu de l'ensemble de données. Nous souhaitons à présent étendre le concept d'identification par un ensemble de données à celui d'**identification asymptotique**. Tout d'abord, observons que dans le Chapitre 2 l'identification était définie en termes de la fonction somme-des-carrés, utilisée pour définir l'estimateur NLS. Puisque nous ne souhaitons pas nous restreindre en permanence au modèle de régression non linéaire, nous ferons simplement référence pour l'instant à un estimateur  $\hat{\theta}$ , sans se soucier de son origine. Naturellement, par "estimateur" nous signifions une *suite* de variables aléatoires,  $\{\hat{\theta}^n\}_{n=m}^{\infty}$ , comme celle discutée dans le Chapitre 4, pour laquelle les éléments de la suite prennent leurs valeurs dans l'espace paramétrique  $\Theta$ . L'élément  $n$  de la suite est une fonction d'un échantillon de taille  $n$ . Formellement, nous pouvons écrire

$$\hat{\theta}^n(\mathbf{y}^n) \in \Theta,$$

où l'exposant  $n$  désigne un échantillon entier de  $n$  observations. Cependant, pour alléger la notation, nous oublierons en général les exposants  $n$  à moins qu'il soit important d'explicitement la dépendance à la taille d'échantillon.

La distinction entre un **estimateur** et une **estimation** n'est pas toujours exposée clairement en économétrie, mais elle peut parfois être utile. La distinction est identique à celle qui existe entre une variable aléatoire (l'*estimateur*) et une réalisation de cette variable aléatoire (l'*estimation*). Ainsi, l'estimateur  $\hat{\theta}^n(\mathbf{y}^n)$  est une *fonction* de l'échantillon aléatoire  $\mathbf{y}^n$ , tandis que l'estimation  $\hat{\theta}$  pour un échantillon donné  $\mathbf{y}$  est la *valeur* de l'estimateur lorsqu'il est évalué en  $\mathbf{y}$ . La notation  $\hat{\theta}$  confond un estimateur et une estimation. Ceci peut dans certaines situations être malvenu, mais les quelques désagréments sont habituellement compensés par la simplicité et la généralité de la notation "chapeau". Nous ferons une distinction explicite lorsque cela sera important.

Le problème de l'identification d'un vecteur de paramètres  $\theta$  par un estimateur  $\hat{\theta}$  est relié au fait que  $\hat{\theta}(\mathbf{y})$  soit déterminé de manière *unique* ou pas pour tout échantillon quelconque  $\mathbf{y}$ , ou plus précisément au fait qu'il puisse exister en tant qu'élément de l'espace paramétrique  $\Theta$ . Comme nous l'avons vu dans la Section 2.3, pour un échantillon donné  $\mathbf{y}$ , la fonction somme-des-carrés  $SSR(\beta)$  peut ne pas atteindre un minimum global quel que soit le



vecteur paramétrique fini  $\beta$ , peut atteindre un minimum global en une valeur interdite telle que  $\beta_2 = 0$  pour le modèle (5.03), ou peut atteindre un minimum globale en plusieurs vecteurs de paramètres  $\beta$ . Dans n'importe lequel de ces cas, l'échantillon  $\mathbf{y}$  n'identifie pas le vecteur de paramètres  $\beta$ . Une simple extension de la définition de l'identification utilisée dans le Chapitre 2 peut alors éliminer ce genre de configuration. Nous avons:

*Définition 5.1.*

Le modèle paramétrisé  $(\mathbb{M}, \theta)$  est identifié par l'échantillon  $\mathbf{y}$  et par l'estimateur  $\hat{\theta}^n$  si  $\hat{\theta}^n(\mathbf{y})$  existe et est unique.

Notons que cette définition s'applique séparément à chaque réalisation possible de l'échantillon  $\mathbf{y}$ , de sorte qu'elle définit une propriété de cet échantillon plutôt que de l'estimateur  $\hat{\theta}^n(\mathbf{y}^n)$ . Ce n'est pas le cas pour le concept de l'**identification asymptotique**, qui est une propriété uniquement rattachée au modèle paramétrisé  $(\mathbb{M}, \theta)$ .

*Définition 5.2.*

Un modèle paramétrisé  $(\mathbb{M}, \theta)$  est dit identifié asymptotiquement si pour n'importe quels  $\theta^1, \theta^2 \in \Theta$  avec  $\theta^1 \neq \theta^2$  il existe une certaine suite de fonctions  $\{Q_n\}$  telle que

$$\text{plim}_{n \rightarrow \infty} Q_n(\mathbf{y}^n) \neq \text{plim}_{n \rightarrow \infty} Q_n(\mathbf{y}^n), \quad (5.04)$$

ou au moins une des limites en probabilité existe et est une constante finie.

La notation

$$\text{plim}_{n \rightarrow \infty} \quad \text{pour } j = 1, 2$$

signifie naturellement que la limite en probabilité est calculée au moyen des DGP caractérisés par des vecteurs paramétriques  $\theta^j$ . La définition étend l'idée qu'un seul vecteur de paramètres peut être associé à un DGP donné et nécessite que deux DGP quelconques caractérisés par des paramètres différents doivent être différents non seulement en échantillon fini mais aussi asymptotiquement. Ceci signifie qu'il est toujours possible de trouver une suite  $Q$  qui discrimine les deux asymptotiquement. Si aucune suite  $Q$  satisfaisant (5.04) n'existait, les propriétés à la limite de n'importe quelle statistique, estimateur ou statistique de test, seraient identiques sous les deux DGP, et nous devrions dans ce cas considérer les DGP comme équivalents asymptotiquement. La Définition 5.2 exclut de façon spécifique la possibilité que des DGP associés à des vecteurs de paramètres différents soient asymptotiquement équivalents dans ce sens.

Le choix le plus évident pour la suite des fonctions  $Q$  dans (5.04) est simplement la  $i^{\text{ième}}$  composante d'un estimateur  $\hat{\theta}$  des paramètres du modèle ayant de "bonnes propriétés", (à supposer qu'un tel estimateur existe), où

naturellement  $i$  est choisi de telle sorte que  $\theta_i^1 \neq \theta_i^2$ . Si l'estimateur possède effectivement de bonnes propriétés, nous pourrions nous attendre à ce que

$$\text{plim}_{n \rightarrow \infty}(\hat{\theta}_i) \neq \text{plim}_{n \rightarrow \infty}(\hat{\theta}_i),$$

et l'estimateur  $\hat{\theta}_i$  discrimine par conséquent  $\theta^1$  et  $\theta^2$  asymptotiquement. Ainsi, l'idée de l'identification asymptotique d'un modèle est à l'évidence étroitement reliée à la *possibilité* de trouver un estimateur pour les paramètres du modèle ayant de bonnes propriétés. Si un modèle n'est pas asymptotiquement identifié, il existe alors au moins deux DGP du modèle, caractérisés par des paramètres différents, tels qu'il *n'existe aucun* estimateur capable de les distinguer asymptotiquement.

L'exemple suivant illustre le cas d'un modèle identifié avec des échantillons finis pour tout ensemble de données, mais néanmoins non identifié asymptotiquement. Il s'agit de l'exemple (4.14) de la tendance temporelle, mais dont la variable a été inversée:

$$y_t = \alpha \frac{1}{t} + u_t, \quad (5.05)$$

avec les aléas  $u_t$  distribués suivant une NID(0,  $\sigma^2$ ) pour un échantillon de taille  $n$ . L'estimateur NLS de  $\alpha$  pour la taille d'échantillon  $n$  est

$$\hat{\alpha}^n = \left( \sum_{t=1}^n t^{-2} \right)^{-1} \left( \sum_{t=1}^n t^{-1} y_t \right).$$

Il s'agit simplement de l'estimateur OLS, puisque (5.05) est une régression linéaire, ce qui garantit que  $\hat{\alpha}^n$  est l'unique valeur donnant un minimum global à la fonction somme-des-carrés. C'est-à-dire que le paramètre  $\alpha$  est identifié par *n'importe quel* ensemble de données  $\mathbf{y}^n$ . Si le véritable DGP est donné par (5.05) avec  $\alpha = \alpha_0$ , alors nous trouvons de manière habituelle que

$$\hat{\alpha}^n = \alpha_0 + \left( \sum_{t=1}^n t^{-2} \right)^{-1} \left( \sum_{t=1}^n t^{-1} u_t \right). \quad (5.06)$$

Dans un modèle ordinaire de régression, la matrice  $\mathbf{X}^\top \mathbf{X}$  est  $O(n)$ . Mais ici, la matrice  $\mathbf{X}^\top \mathbf{X}$  est la quantité scalaire  $\sum_{t=1}^n t^{-2}$ , et la série  $\sum_{t=1}^n t^{-2}$  converge vers la limite  $\pi^2/6$  quand  $n \rightarrow \infty$ .<sup>1</sup> Le facteur aléatoire  $\sum_{t=1}^n t^{-1} u_t$  est normalement distribué avec une espérance nulle, tout comme les  $u_t$ , et une variance égale à  $\sigma^2 \sum_{t=1}^n t^{-2}$ , une grandeur qui tend vers  $\sigma^2 \pi^2/6$ . Ainsi,  $\hat{\alpha}^n$  égale  $\alpha_0$  plus une variable aléatoire normale d'espérance nulle et de variance tendant vers  $6\sigma^2/\pi^2$  quand  $n \rightarrow \infty$ . Cette variance limite *n'est pas* nulle, et

<sup>1</sup> Consulter, par exemple, Abramowitz et Stegun (1965), équation 23.2.24, page 807, ou n'importe quelle discussion de la fonction zeta de Riemann.

la limite en probabilité de l'estimateur  $\hat{\alpha}$  est par conséquent non dégénérée: ce n'est pas une constante déterministe. En fait, la limite en probabilité sera différente pour des valeurs différentes de  $\alpha_0$ , mais la Définition 5.2 requiert des fonctions  $Q$  ayant des limites en probabilité *non stochastiques*.

Il serait très fastidieux de montrer qu'il peut *ne pas exister* une suite  $Q$  satisfaisant les conditions de la Définition 5.2 et capable de distinguer des valeurs différentes du paramètre  $\alpha$  de (5.05). La compréhension intuitive de la non identification asymptotique de (5.05) est beaucoup plus importante qu'une démonstration formelle de la proposition précédente, puisque les lecteurs peuvent très bien considérer que la nécessité d'une limite en probabilité non stochastique n'est qu'un subterfuge. Le point clé est simplement que pour un modèle de régression linéaire de la forme (5.01) dans lequel  $\mathbf{X}^\top \mathbf{X} = O(n)$  quand  $n \rightarrow \infty$ , la matrice de covariance des paramètres estimés tend vers zéro quand  $n \rightarrow \infty$ . Ceci signifie que, lorsque la taille de l'échantillon augmente, la **précision**<sup>2</sup> de l'estimateur OLS devient arbitrairement grande. C'est également le cas pour l'estimateur NLS, comme nous le verrons dans la suite du chapitre. Par contraste, comme nous l'avons vu à partir de (5.06), la précision de  $\hat{\alpha}$  dans le modèle (5.05) tend vers une limite finie, quelle que soit la taille de l'échantillon. C'est justement pour éliminer des modèles tels que (5.05) que la Définition 5.2 contient les conditions nécessaires qui y figurent.

Un modèle paramétrisé peut être asymptotiquement identifié mais non asymptotiquement identifié par un estimateur particulier. N'importe quel estimateur satisfaisant devrait être capable, comme la suite  $Q$  de la Définition 5.2, de discriminer des DGP caractérisés par des vecteurs paramétriques différents s'il a pour but d'estimer le vecteur paramétrique. Naturellement, ceci n'est pas possible si le modèle n'est pas lui-même asymptotiquement identifié, mais s'il l'est, un estimateur satisfaisant doit être capable d'identifier les paramètres du modèle asymptotiquement. La propriété requise pour un tel estimateur est celle de la **convergence**, dont nous avons discuté dans la Section 4.5. De façon formelle:

*Définition 5.3.*

Un estimateur  $\hat{\theta} \equiv \{\hat{\theta}^n\}$  estime de façon convergente les paramètres d'un modèle paramétrisé  $(\mathbb{M}, \theta)$ , ou converge, si pour tout  $\mu_0 \in \mathbb{M}$ ,

$$\text{plim}_0(\hat{\theta}^n) = \theta_0. \quad (5.07)$$

La notation “ $\text{plim}_0$ ” signifie simplement que nous calculons la limite en probabilité sous le DGP  $\mu_0$ , caractérisé par  $\theta_0$ .

<sup>2</sup> La **précision** d'une variable aléatoire est simplement la réciproque de sa variance, et la **matrice de précision** d'une variable aléatoire vectoriel est l'inverse de sa matrice de covariance. En dépit de la simplicité de la relation entre les deux concepts, il est parfois plus intuitif de raisonner en terme de précision plutôt qu'en terme de variance.

Un estimateur convergent fournit évidemment une suite  $Q$  pour la Définition 5.2, puisque pour  $\mu_1$  et  $\mu_2 \in \mathbb{M}$  tels que  $\boldsymbol{\theta}_1 \equiv \boldsymbol{\theta}(\mu_1) \neq \boldsymbol{\theta}(\mu_2) \equiv \boldsymbol{\theta}_2$ , il découle immédiatement de (5.07) que

$$\text{plim}_1(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 = \text{plim}_2(\hat{\boldsymbol{\theta}}),$$

comme le demandait la définition. Ainsi, n'importe quel modèle paramétrisé pour lequel un estimateur convergent existe est, *a fortiori*, identifié asymptotiquement. Cependant, tous les estimateurs concevables d'un modèle identifié asymptotiquement ne parviennent pas à l'identifier asymptotiquement. Dans la pratique, ceci survient rarement. L'on est rarement confronté à deux procédures d'estimation fiables, l'une identifiant asymptotiquement les paramètres d'un modèle paramétrisé, et l'autre pas. Mais une exception curieuse et importante à cette remarque est que l'estimateur NLS n'identifie pas la variance des aléas  $\sigma^2$ , puisque la fonction somme-des-carrés, qui définit l'estimateur NLS, ne dépend pas de  $\sigma^2$ . Cela a finalement peu d'importance, puisque nous pouvons quand même estimer  $\sigma^2$ , mais cela distingue l'estimation NLS des autres méthodes, tel le maximum de vraisemblance, qui identifient la variance des aléas.

Dans la prochaine section, nous porterons notre attention sur l'estimateur NLS et démontrerons que dans le cas d'un modèle identifié asymptotiquement par cet estimateur, il est convergent.

### 5.3 CONVERGENCE DE L'ESTIMATEUR NLS

Un “modèle de régression non linéaire” univarié a jusqu'à présent été écrit

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad (5.08)$$

où  $\mathbf{y}$ ,  $\mathbf{x}(\boldsymbol{\beta})$ , et  $\mathbf{u}$  sont des vecteurs de dimension  $n$  pour une taille d'échantillon  $n$  donnée. Les paramètres du modèle sont par conséquent  $\boldsymbol{\beta}$  et soit  $\sigma$  soit  $\sigma^2$ . La fonction de régression  $x_t(\boldsymbol{\beta})$ , qui est le  $i^{\text{ième}}$  élément de  $\mathbf{x}(\boldsymbol{\beta})$ , dépendra en général du vecteur ligne des variables  $\mathbf{Z}_t$ . La spécification du vecteur des aléas  $\mathbf{u}$  n'est pas complète, puisque la distribution des  $u_t$  n'a pas été spécifiée. Ainsi, pour un échantillon de taille  $n$ , le modèle  $\mathbb{M}$  décrit par (5.08) est l'ensemble de tous les DGP qui génèrent des échantillons  $\mathbf{y}$  de taille  $n$  tels que l'espérance de  $y_t$  conditionnellement à un ensemble d'information  $\Omega_t$  quelconque qui comprend  $\mathbf{Z}_t$ , est  $x_t(\boldsymbol{\beta})$  pour un certain vecteur paramétrique  $\boldsymbol{\beta} \in \mathbb{R}^k$ , et tels que les différences  $y_t - x_t(\boldsymbol{\beta})$  sont des aléas indépendamment distribués et de variance commune  $\sigma^2$ , habituellement inconnue.

Il sera commode de généraliser quelque peu cette spécification des DGP dans  $\mathbb{M}$ , afin de pouvoir traiter les **modèles dynamiques**, c'est-à-dire les modèles comprenant des **variables dépendantes retardées**. Par conséquent,

nous reconnaissons explicitement la possibilité que la fonction de régression  $x_t(\beta)$  puisse dépendre d'un nombre arbitraire mais borné de retards de la variable dépendante elle-même. Ainsi,  $x_t$  peut dépendre de  $y_{t-1}, y_{t-2}, \dots, y_{t-l}$ , où  $l$  est un entier positif fixé qui ne dépend pas de la taille d'échantillon. Quand le modèle utilise des données temporelles,  $x_t(\beta)$  désignera l'espérance de  $y_t$  conditionnellement à un ensemble d'information qui comprend le passé entier de la variable dépendante, que nous pouvons noter  $\{y_s\}_{s=1}^{t-1}$ , mais également l'histoire entière des variables exogènes jusqu'à la période  $t$  comprise, à savoir,  $\{\mathbf{Z}_t\}_{s=1}^t$ . Les propriétés du vecteur de perturbations  $\mathbf{u}$  restent inchangées.

Pour que la théorie asymptotique soit applicable, nous devons ensuite fournir une règle pour le développement de (5.08) à des échantillons de taille arbitrairement grande. Pour des modèles qui ne sont pas dynamiques (et parmi eux les modèles estimés avec des données en coupe transversale, naturellement), tels qu'il n'existe aucune tendance temporelle ou variables dépendantes retardées dans les fonctions de régression  $x_t$ , rien n'empêche la simple utilisation du concept de régresseurs "fixes en échantillons répétés" dont nous avons discuté dans la Section 4.4. Typiquement, nous ne considérons que des tailles d'échantillon qui sont des entiers multiples de la taille d'échantillon réelle  $m$  et supposons alors que  $x_{Nm+t}(\beta) = x_t(\beta)$  pour  $N > 1$ . Cette hypothèse rend les propriétés asymptotiques des modèles non dynamiques extrêmement simples par rapport à celles des modèles dynamiques.<sup>3</sup>

Certains économètres opposeraient l'idée que la solution précédente est trop simpliste lorsque l'on travaille avec des données temporelles et préféreraient une règle comme celle qui suit. Les variables  $\mathbf{Z}_t$  qui apparaissent dans les fonctions de régression manifesteront habituellement elles-mêmes les mêmes régularités que les séries temporelles et seront susceptibles d'être modélisées comme l'un des processus stochastiques standards utilisés dans l'analyse des séries temporelles; nous discuterons avec plus de détails de ces processus standards dans le Chapitre 10. Afin d'étendre le DGP (5.08), les valeurs hors échantillon pour les  $\mathbf{Z}_t$  devraient être elles-mêmes considérées comme aléatoires, étant générées par des processus appropriés. L'introduction de cet aléa supplémentaire complique quelque peu l'analyse asymptotique, mais pas énormément, puisque nous supposerions malgré tout que les processus stochastiques générant les  $\mathbf{Z}_t$  sont indépendants du processus stochastique qui génère le vecteur de perturbations  $\mathbf{u}$ .

Dans le cas des modèles dynamiques, il est incontestable que la seconde méthode d'extension des DGP, basée sur la recherche de représentations standards des séries temporelles pour toutes les variables comprises dans les

<sup>3</sup> En effet, même pour des modèles dynamiques *linéaires*, il n'est pas trivial de montrer que les moindres carrés fournissent des estimations convergentes, asymptotiquement normales. La référence classique sur ce sujet est Mann et Wald (1943).

fonctions de régression, est intuitivement plus satisfaisante que le concept de régresseurs fixes en échantillons répétés. L'équation de régression (5.08) doit alors s'interpréter comme une équation de différence stochastique, définissant le processus stochastique qui génère le vecteur des observations sur la variable dépendante,  $\mathbf{y}$ , en terme des réalisations de  $\mathbf{Z}$  et des **innovations**  $\mathbf{u}$ . Nous supposerons toujours que ces dernières sont des bruits blancs, c'est-à-dire qu'elles ont la propriété  $\mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  pour n'importe quelle taille d'échantillon  $n$ . De plus, nous supposerons que le processus stochastique qui génère les innovations est indépendant des processus stochastiques qui génèrent les  $\mathbf{Z}_t$  et que ces derniers sont *identiques* pour tous les DGP du modèle. Ainsi, les véritables valeurs des paramètres du modèle n'influencent pas les processus qui génèrent toutes les variables autres que la variable dépendante.

Le problème qui consiste à connaître la relation entre les variables dont dépendent les fonctions de régression et la (les) variable(s) dépendante(s), a été un sujet majeur de la recherche en économétrie. Les hypothèses posées concernant les  $\mathbf{Z}_t$  impliquent qu'elles sont **exogènes** ou, pour utiliser la terminologie de Engle, Hendry, et Richard (1983), **strictement exogènes**. De nombreuses définitions de l'exogénéité sont disponibles, comme le montre la lecture de cet article. Nous considérerons le sujet en détails lorsque nous serons amenés à discuter des modèles d'équations simultanées dans le Chapitre 18.

Avec le préambule précédent, nous sommes prêts à discuter de la convergence de l'estimateur NLS. Pour cela, autant que pour le traitement de l'identification asymptotique, nous étudions les propriétés de la fonction somme-des-carrés quand la taille d'échantillon  $n$  tend vers l'infini. Puisque cette fonction ne dépend pas de  $\sigma^2$ , nous simplifierons notre tâche en supposant que, bien que tout DGP  $\mu$  dans le modèle  $\mathbb{M}$  considéré soit tel que toutes les observations sont caractérisées par une variance des aléas unique (inconnue)  $\sigma^2$ , ce paramètre n'est pas défini par l'application définissante des paramètres, que nous noterons  $\beta$  plutôt que  $\theta$  pour des raisons évidentes. Ainsi en excluant  $\sigma^2$  de la liste des paramètres du modèle, nous pouvons nous concentrer sur la capacité de l'estimateur NLS à identifier les autres paramètres du modèle, à savoir ceux de la fonction de régression.

Nous rendrons explicite la dépendance de la fonction somme-des-carrés à la taille d'échantillon  $n$  et à l'échantillon  $\mathbf{y}$ :

$$SSR^n(\mathbf{y}, \beta) \equiv \sum_{t=1}^n (y_t - x_t(\beta))^2. \quad (5.09)$$

Cette fonction est la somme de  $n$  termes non négatifs qui ne tendront pas en général vers zéro quand  $n \rightarrow \infty$ , de sorte que leur somme tendra en général vers l'infini avec  $n$ . Puisque l'infini n'est pas traditionnellement une limite intéressante, nous préférons travailler avec la moyenne de ces termes plutôt qu'avec leur somme. Ainsi, nous définissons

$$ssr^n(\mathbf{y}, \beta) \equiv n^{-1} SSR^n(\mathbf{y}, \beta). \quad (5.10)$$

Puisque la fonction  $ssr^n$  est définie comme une moyenne, nous pouvons espérer être capables de lui appliquer une loi des grands nombres. Si tel est le cas, nous pouvons alors poser la définition suivante:

$$\overline{ssr}(\beta, \mu) \equiv \text{plim}_{\mu} ssr^n(\mathbf{y}, \beta) = \lim_{n \rightarrow \infty} E_{\mu}(ssr^n(\mathbf{y}, \beta)), \quad (5.11)$$

où  $\text{plim}_{\mu}$  et  $E_{\mu}$  indiquent que nous calculons la limite en probabilité ou l'espérance sous le DGP  $\mu$ .

Il est possible d'exprimer la question de l'identification asymptotique d'un modèle de régression non linéaire en terme de la fonction limite  $\overline{ssr}$ , si elle existe, tout comme la question de l'identification ordinaire peut s'exprimer en termes de  $SSR^n$  ou, de façon équivalente,  $ssr^n$ . La fonction  $\overline{ssr}$  existe-t-elle en général? Au niveau de généralité auquel nous nous situons jusqu'à présent, c'est-à-dire celui des modèles dynamiques en général, la réponse est négative. Nous discuterons de ce point plus en détail par la suite. Cependant, à moins que  $\overline{ssr}$  n'existe, la discussion de ce chapitre est inapplicable. Il existe des modèles pour lesquels  $\overline{ssr}$  n'existe pas, mais qui sont néanmoins asymptotiquement identifiées par l'estimateur NLS; le Chapitre 20 apporte certains exemples. Cependant, pour de tels modèles l'estimateur NLS n'aura pas les propriétés standards de normalité asymptotique et une convergence au taux  $n^{1/2}$  que nous démontrons dans la prochaine section.

La propriété de  $\overline{ssr}$  qui implique la convergence de l'estimateur NLS est la suivante. Soit  $\beta_0$  et  $\sigma_0$  les valeurs de  $\beta$  et  $\sigma$  sous le DGP  $\mu_0$  qui a réellement généré les données. Nous pouvons alors montrer, sous des conditions de régularité adéquates,

$$\overline{ssr}(\beta_0, \mu_0) < \overline{ssr}(\beta, \mu_0) \quad \text{pour tout } \beta \neq \beta_0. \quad (5.12)$$

Autrement dit, la limite en moyenne des résidus au carré est minimisée lorsque les résidus sont évalués avec le véritable vecteur paramétrique  $\beta_0$ . Pourquoi cela implique-t-il la convergence? Sans entrer dans des détails techniques, cela peut se comprendre si nous acceptons que la limite des estimateurs NLS en échantillon fini  $\hat{\beta}^n$ , définis de façon à minimiser  $ssr^n$ , est la valeur de  $\beta$  qui minimise la fonction limite  $\overline{ssr}$ . Ainsi, cette valeur, d'après (5.12), est simplement la véritable valeur  $\beta_0$ .

Bien que plausible, cet argument est faussement simple. Lorsque nous établirons un argument comparable dans le Chapitre 8, dans le contexte de l'estimation par maximum de vraisemblance, nous serons plus prudents, sans toutefois être pleinement rigoureux. Pour l'instant, nous nous contentons de présenter un théorème dans lequel nous supposons suffisamment de régularité pour que le passage de (5.12) à la convergence de l'estimateur NLS soit justifié. Nous discuterons ensuite, dans certains cas pratiques importants, de l'existence et des conditions de l'existence de  $\overline{ssr}$ , et de la validité et des conditions de validité de (5.12).

*Théorème 5.1. Théorème de Convergence des Moindres Carrés non Linéaires.*

Supposons que

- (i) le modèle de régression non linéaire (5.08), considéré comme un modèle paramétrisé  $(\mathbb{M}, \beta)$ , avec un espace paramétrique  $\Theta$ , est asymptotiquement identifié par la fonction  $\overline{ssr}$ . Ainsi, pour tout  $\mu_0 \in \mathbb{M}$ ,

$$\overline{ssr}(\beta_0, \mu_0) \neq \overline{ssr}(\beta, \mu_0) \quad (5.13)$$

pour tout  $\beta \in \Theta$  tel que  $\beta \neq \beta_0$ ;

- (ii) la suite  $\{n^{-1} \sum_{t=1}^n x_t(\beta) u_t\}$  satisfait la condition WULLN de la Définition 4.17 avec une limite en probabilité nulle, pour chaque  $\mu_0 \in \mathbb{M}$  et pour tout  $\beta \in \Theta$ ; et
- (iii) la limite en probabilité de la suite  $\{n^{-1} \sum_{t=1}^n x_t(\beta) x_t(\beta')\}$ , pour n'importe quel  $\beta' \in \Theta$ , est finie, continue en  $\beta$  et  $\beta'$ , non stochastique, et uniforme par rapport à  $\beta$  et  $\beta'$ ,

alors l'estimateur NLS  $\hat{\beta}$  converge vers les paramètres  $\beta_0$ .

Nous ne démontrerons pas ce théorème mais tenterons de faire comprendre intuitivement à quoi servent les diverses conditions de régularité du théorème. Tout d'abord, observons que dans la condition (i) du théorème, nous demandons que le modèle soit asymptotiquement identifié par  $\overline{ssr}$ . Cette fonction n'est pas un estimateur, contrairement aux autres fonctions jouant le rôle de la suite  $Q$  dans la Définition 5.2 jusqu'à présent, mais la fonction qui définit l'estimateur NLS asymptotiquement, de sorte qu'il est commode d'exprimer la condition d'identification asymptotique en terme de cette fonction. La condition (5.13) est légèrement plus compliquée que (5.04). La raison est que  $\overline{ssr}(\beta, \mu_0)$ , en tant que fonction scalaire, prendra la même valeur pour des valeurs différentes de  $\beta \neq \beta_0$ . Mais il faut seulement que ces valeurs soient toutes différentes de  $\overline{ssr}(\beta_0, \mu_0)$ .

Examinons à présent plus attentivement  $\overline{ssr}$  et l'inégalité (5.12). A partir de (5.08), (5.09), et (5.10) nous avons

$$\begin{aligned} ssr^n(\mathbf{y}, \beta) &= \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta) + u_t)^2 \\ &= \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta))^2 + \frac{2}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta)) u_t + \frac{1}{n} \sum_{t=1}^n u_t^2. \end{aligned} \quad (5.14)$$

Le dernier terme de la dernière expression ici est le plus facile à traiter. Puisque les variables aléatoires  $u_t$  sont i.i.d. sous  $\mu_0$ , le Théorème 4.3, la plus simple des lois des grands nombres, s'applique immédiatement et donne

$$\text{plim}_0 \left( \frac{1}{n} \sum_{t=1}^n u_t^2 \right) = E(u_t^2) = \sigma_0^2.$$



Examinons ensuite le second terme dans la dernière expression de (5.14). Tout phénomène aléatoire dans les fonctions de régression  $x_t$  doit provenir soit de la présence de variables dépendantes retardées soit d'un phénomène aléatoire dans les  $\mathbf{Z}_t$ , les autres variables dont peuvent dépendre les fonctions de régression, qui serait alors indépendant des perturbations  $u_t$ . Des variables dépendantes retardées à la période  $t$  ne peuvent dépendre que des perturbations contenues dans la suite  $\{u_s\}_{s=1}^{t-1}$ , et celles-ci sont naturellement indépendantes de  $u_t$  lui-même. Ainsi, dans toutes circonstances les deux facteurs dans chaque terme de la somme

$$\frac{2}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta)) u_t = \frac{2}{n} \sum_{t=1}^n x_t(\beta_0) u_t - \frac{2}{n} \sum_{t=1}^n x_t(\beta) u_t \quad (5.15)$$

sont indépendants, de sorte que chaque terme de cette somme a une espérance nulle, puisque  $u_t$  a une espérance nulle. Cependant, les termes successifs ne sont pas nécessairement mutuellement indépendants, puisque la présence des variables dépendantes retardées dans  $x_t(\beta_0) - x_t(\beta)$  mènerait à une possible corrélation de cette expression avec les termes indicés par  $t-1, \dots, t-i$  de la somme (5.15), et la plupart des représentations des  $\mathbf{Z}_t$  comme séries temporelles mènera également à de telles corrélations. Ainsi, si nous devons utiliser une loi des grands nombres afin de conclure que la limite en probabilité de (5.15) est nulle, nous devons expliciter les hypothèses suffisantes qui garantissent qu'une telle loi des grands nombres s'applique. Il est nécessaire de pouvoir appliquer une loi uniforme des grands nombres à

$$\frac{1}{n} \sum_{t=1}^n x_t(\beta) u_t \quad (5.16)$$

pour tout  $\beta \in \Theta$ . C'est la raison pour laquelle la condition (ii) du Théorème 5.1 a été imposée.

Pour le premier terme dans la dernière expression de (5.14) nous souhaitons pouvoir appliquer une loi uniforme des grands nombres à

$$\frac{1}{n} \sum_{t=1}^n x_t(\beta) x_t(\beta') \quad (5.17)$$

pour des valeurs arbitraires de  $\beta, \beta' \in \Theta$ , et ceci justifie la condition (iii) du théorème.

Sous les conditions du théorème, nous obtenons ensuite le résultat

$$\overline{ssr}(\beta, \mu_0) = \sigma_0^2 + \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta))^2 \right). \quad (5.18)$$

Il est immédiatement évident à partir de (5.18) que  $\overline{ssr}(\beta, \mu_0)$  est minimisée lorsque  $\beta = \beta_0$ , ce qui rend (5.12) valable avec une inégalité faible. L'inégalité

stricte requise pour la convergence de l'estimateur NLS est fournie par la condition (i) du théorème, la condition d'identification asymptotique.

Peut-on trouver des conditions suffisantes aisément compréhensibles pour les conditions du Théorème 5.1? Oui, mais malheureusement, les conditions aisément interprétables tendent à être trop restrictives. L'une des hypothèses les plus simples est que les fonctions de régression  $x_t(\beta)$  sont indépendantes et uniformément bornées. Ceci permet l'usage de la loi des grands nombres du Théorème 4.4, de laquelle nous pouvons conclure que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta)) u_t \right) = 0$$

et également que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta))^2 \right) \quad (5.19)$$

existe et est une grandeur non négative, non stochastique. Si le modèle est asymptotiquement identifié, cette grandeur sera strictement positive pour tout  $\beta \neq \beta_0$ .

L'hypothèse d'indépendance est naturellement souvent beaucoup trop forte. Plus généralement, nous aimerions considérer le cas d'une fonction de régression  $x_t(\beta)$  qui dépend seulement de variables non aléatoires et d'un nombre fini de variables dépendantes retardées:

$$x_t(\beta) = x_t(\mathbf{Z}_t, y_{t-1}, \dots, y_{t-i}; \beta). \quad (5.20)$$

Malheureusement, la forme (5.20) *n'est pas* en général telle qu'une loi des grands nombres puisse s'appliquer à (5.16) et (5.17). Le cas le plus flagrant est celui d'un **processus explosif**, dont un exemple particulièrement simple est fourni par le DGP

$$y_t = \alpha y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2) \quad (5.21)$$

pour tout  $\alpha$  avec  $|\alpha| > 1$ . Il est aisé de voir pourquoi cette spécification conduit à un processus explosif: supposons que la variance de  $y_1$  soit  $\sigma_1^2$ , et calculons la variance de  $y_t$ . Nous trouvons que

$$\begin{aligned} \text{Var}(y_t) &= \text{Var}(\alpha y_{t-1} + u_t) \\ &= \alpha^2 \text{Var}(y_{t-1}) + \sigma^2 \\ &= \alpha^4 \text{Var}(y_{t-2}) + \sigma^2(1 + \alpha^2) \\ &= \alpha^{2(t-1)} \sigma_1^2 + \sigma^2(\alpha^2 - 1)^{-1}(\alpha^{2(t-1)} - 1), \end{aligned} \quad (5.22)$$

où la dernière ligne dans (5.22) est obtenue par la substitution répétée du résultat contenu dans la première ligne. Nous voyons immédiatement que,

puisque  $|\alpha| > 1$ , la variance de  $y_t$  tend vers l'infini avec  $t$ . Le terme qui correspond à  $x_t(\beta)u_t$  pour la fonction de régression  $\alpha y_{t-1}$  de (5.21) est  $\alpha y_{t-1}u_t$ , et nous voyons que la variance de ce terme tend également vers l'infini avec  $t$ . Ainsi, aucune loi des grands nombres ne peut s'appliquer en général à (5.16).

Les économètres veillent habituellement à ce que les fonctions de régression qu'ils utilisent ne donnent pas naissance à des processus explosifs tels que celui considéré. Si nous imposons que  $|\alpha| < 1$  dans (5.21), nous obtenons un processus qui n'est pas explosif.<sup>4</sup> Afin de pouvoir gérer ce cas, et plus généralement une fonction de régression (5.20) quand elle ne conduit pas à un processus explosif, la loi des grands nombres la plus utile est celle de la martingale, Théorème 4.6. Notons tout d'abord que ce théorème peut être appliqué directement aux termes  $x_t(\beta)u_t$ , puisque l'espérance de  $x_t(\beta)u_t$ , conditionnellement à  $\{x_s(\beta)u_s\}_{s=1}^{t-1}$ , est nulle, parce que  $u_t$  est indépendant à la fois de  $u_s$  et de  $x_s(\beta)$  pour tout  $s \leq t$ . Ainsi, la seule contrainte supplémentaire du théorème est très faible et peut être satisfaite en imposant que les espérances des  $x_t(\beta)$  soient uniformément bornées.

Reste la question de notre certitude sur l'existence de l'expression (5.19) et sur le fait qu'elle soit non stochastique. C'est une question à laquelle on ne peut répondre que si la fonction de régression et le DGP ont été spécifiés en détail. Nous adopterons donc la position que (5.19) existe et est non stochastique si le processus défini par la fonction de régression (5.20) n'est pas explosif. Ainsi, quand nous dirons qu'un processus n'est pas explosif, nous signifierons que (5.19) existe, est finie et non stochastique. Par ce biais, nous pouvons considérer des modèles de régression non linéaire avec des fonctions de régression comme (5.20) individuellement afin de déterminer s'ils sont explosifs.

Considérons par exemple le modèle simple (5.21), mais avec  $|\alpha| < 1$ . Pour cette spécification, (5.19) devient

$$(\alpha_0 - \alpha)^2 \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n y_{t-1}^2 \right). \quad (5.23)$$

Pour nos propos le facteur  $(\alpha_0 - \alpha)^2$  est non pertinent, et il suffit d'étudier la limite en probabilité. Malheureusement, ceci n'est pas particulièrement facile sans faire appel à des propriétés des processus stochastiques que nous n'avons pas exposées jusqu'à présent et n'exposerons pas dans cet ouvrage.<sup>5</sup> Mais nous verrons dans le Chapitre 10 que la suite  $\{y_t\}$  générée par (5.21) est ce que l'on appelle un processus autorégressif d'ordre 1, ou processus AR(1), et que pour  $|\alpha| < 1$  il est stationnaire et ergodique. Les mêmes propriétés sont valables par

<sup>4</sup> De tels processus seront discutés dans le Chapitre 10 en connexion avec notre discussion de l'autocorrélation.

<sup>5</sup> Consulter Lamperti (1977) pour une discussion plus générale des processus stochastiques à un niveau avancé.

conséquence pour la suite  $\{y_t^2\}$ . Donc, nous pouvons appliquer le théorème ergodique, Théorème 4.5, afin d'obtenir le résultat désiré que le processus (5.21) avec  $|\alpha| < 1$  n'est pas explosif. L'estimateur des moindres carrés non linéaires du paramètre  $\alpha$  dans (5.21), qui est ici simplement l'estimateur OLS, naturellement, est par conséquent convergent. Ceci provient du Théorème 5.1, puisque l'uniformité de la convergence requise peut s'interpréter comme une conséquence de la structure de (5.23) en tant que produit d'un facteur ne dépendant que du paramètre  $\alpha$  et d'un facteur ne dépendant que du processus aléatoire  $\{y_t\}$ .

Si la discussion précédente semble quelque peu désinvolte sur les processus explosifs, il est malheureusement nécessaire dans l'état actuel de connaissances qu'il en soit ainsi. Il est souvent extrêmement difficile de savoir, même si l'on dispose d'un temps de calcul illimité pour essayer une multitude de simulations variées, si le processus stochastique généré par une certaine fonction de régression donnée de la forme (5.20) est explosif ou non. Les lecteurs intéressés sont encouragés à consulter White (1984) pour se forger une idée de la complexité mathématique employée. En dehors du contexte des processus standards de séries temporelles (qui ne contiennent aucune variable autre que la variable dépendante elle-même; consulter le Chapitre 10) nous pouvons dire très peu de choses en général. Les économètres praticiens peuvent être pardonnés de leur sentiment que la complexité mathématique est inutile, puisque la clé du problème n'est pas empirique mais relative à la meilleure manière de modéliser les données. Nous discuterons d'un certain nombre de problèmes reliés aux processus non stationnaires dans le Chapitre 20.

#### 5.4 NORMALITÉ ASYMPTOTIQUE DE L'ESTIMATEUR NLS

Dans cette section, nous discutons de la normalité asymptotique de l'estimateur des moindres carrés non linéaires. Pour cela, nous demanderons un peu plus de régularité que cela était nécessaire pour la convergence, comme nous allons le voir. Tout d'abord, une définition formelle de la normalité asymptotique:

*Définition 5.4.*

Un estimateur convergent  $\hat{\beta} \equiv \{\hat{\beta}^n\}$  des paramètres du modèle paramétrisé identifié asymptotiquement  $(\mathbb{M}, \beta)$  est asymptotiquement normal si pour chaque DGP  $\mu_0 \in \mathbb{M}$ , la suite des variables aléatoires  $\{n^{1/2}(\hat{\beta}^n - \beta_0)\}$  tend en distribution vers une distribution normale (multivariée), avec une espérance nulle et une matrice de covariance finie.

La différence cruciale entre la propriété de normalité asymptotique et celle de convergence discutée dans la section précédente est le facteur de  $n^{1/2}$ . Ce facteur "amplifie"  $\hat{\beta} - \beta_0$ , qui, si  $\hat{\beta}$  converge vers  $\beta_0$ , tend vers zéro quand

$n$  tend vers l'infini. Ainsi, le produit  $n^{1/2}(\hat{\beta} - \beta_0)$  tend vers un vecteur de variables aléatoires non nulles. La normalité asymptotique, lorsqu'elle est valable, impliquera naturellement la convergence, puisque si  $n^{1/2}(\hat{\beta} - \beta_0)$  est  $O(1)$ , il s'ensuit que  $\hat{\beta} - \beta_0$  doit être  $O(n^{-1/2})$ . Si l'estimateur  $\hat{\beta}$  satisfait la dernière propriété, il est dit **convergent au taux**  $n^{1/2}$ , ce qui signifie que la différence entre l'estimateur et la véritable valeur est proportionnelle à 1 sur  $\sqrt{n}$ . Un estimateur convergent au taux  $n^{1/2}$  doit aussi être faiblement convergent, puisque  $\text{plim}(\hat{\beta} - \beta_0) = \mathbf{0}$ . Cependant, tous les estimateurs ne sont pas convergents au taux  $n^{1/2}$ .

Comme dans la section précédente, nous établirons tout d'abord un théorème qui fournit les conditions suffisantes à la normalité asymptotique de l'estimateur NLS et discuterons ensuite des circonstances dans lesquelles nous pouvons espérer que ces conditions seront satisfaites. Pour commencer, quelques notations. Soit  $\mathbf{X}_t(\beta) \equiv D_\beta x_t(\beta)$  le vecteur ligne des dérivées partielles de la fonction de régression  $x_t(\beta)$ ; alors  $\mathbf{A}_t(\beta) \equiv D_{\beta\beta} x_t(\beta)$  désignera la matrice Hessienne de  $x_t(\beta)$ , et  $\mathbf{H}_t(y_t, \beta) \equiv D_{\beta\beta}(y_t - x_t(\beta))^2$  désignera la matrice Hessienne de la contribution de l'observation  $t$  à la fonction somme-des-carrés. Cette dernière matrice est

$$\mathbf{H}_t(y_t, \beta) = 2\left(\mathbf{X}_t^\top(\beta)\mathbf{X}_t(\beta) - \mathbf{A}_t(\beta)(y_t - x_t(\beta))\right). \quad (5.24)$$

Evidemment, la matrice Hessienne  $\mathbf{A}_t$  de la fonction de régression sera une matrice nulle si la fonction de régression  $x_t$  est linéaire, et  $\mathbf{X}_t(\beta)$  sera simplement  $\mathbf{X}_t$ . Dans ce cas, la matrice  $\mathbf{H}_t(y_t, \beta)$  se simplifiera pour donner  $2(\mathbf{X}_t^\top \mathbf{X}_t)$ , qui est nécessairement semi-définie positive.

*Théorème 5.2. Théorème de Normalité Asymptotique des Moindres Carrés non Linéaires.*

Si le modèle de régression non linéaire (5.08) est asymptotiquement identifié et satisfait les conditions de régularité du Théorème 5.1, de telle sorte que l'estimateur NLS pour le modèle est convergent, et si de plus, pour tout  $\mu_0 \in \mathbb{M}$ ,

- (i) la suite  $\{n^{-1} \sum_{t=1}^n \mathbf{H}_t(y_t, \beta)\}$  satisfait la condition WULLN de la Définition 4.17 pour  $\beta$  dans le voisinage de  $\beta_0$ , et
- (ii) la suite  $\{n^{-1/2} \sum_{t=1}^n \mathbf{X}_t^\top(\beta) u_t\}$  satisfait la condition CLT de la Définition 4.16, et
- (iii) la matrice Hessienne de la fonction limite somme-des-carrés évaluée avec les véritables paramètres,  $D_{\beta\beta} \overline{ssr}(\beta_0, \mu_0)$ , est une matrice définie positive, qui garantit que la condition suffisante du second ordre pour le minimum dans (5.12) est satisfaite,

alors, sous tous les DGP  $\mu_0$  tels que  $\beta_0$  appartient à l'intérieur de l'espace paramétrique  $\Theta$ , l'estimateur NLS  $\hat{\beta}$  est asymptotiquement normal comme dans la Définition 5.4. De plus, si  $\sigma_0^2$  est la variance

des aléas associée à  $\mu_0$ , la **matrice de covariance asymptotique** de  $n^{1/2}(\hat{\beta} - \beta_0)$  est

$$\sigma_0^2 \operatorname{plim}_0 \left( n^{-1} \mathbf{X}_0^\top \mathbf{X}_0 \right)^{-1}. \quad (5.25)$$

Ici,  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$  désigne la matrice de dimension  $n \times k$  avec comme ligne type  $\mathbf{X}_t(\beta_0)$ .

Nous commençons notre discussion de ce théorème à partir de la contrainte que le DGP, que nous noterons  $\mu_0$ , soit tel que  $\beta_0$  appartienne à l'intérieur de l'espace paramétrique  $\Theta$ . Si tel est le cas, alors avec une probabilité arbitrairement proche de l'unité, ce sera également le cas de l'estimateur  $\hat{\beta}$  pour un  $n$  assez grand, puisque nous avons supposé que  $\hat{\beta}$  est convergent. Ceci signifie que  $\hat{\beta}$  doit satisfaire la condition nécessaire du premier ordre pour un minimum intérieur:

$$D_{\beta} \operatorname{ssr}^n(\mathbf{y}, \hat{\beta}) = \mathbf{0}. \quad (5.26)$$

La convergence de  $\hat{\beta}$  signifie qu'il doit être *proche* de  $\beta_0$  si  $n$  est grand. En conséquence, nous opérerons un développement de Taylor à l'ordre un de (5.26) autour de  $\beta_0$ , comme suit:

$$\mathbf{0} = D_{\beta} \operatorname{ssr}^n(\mathbf{y}, \beta_0) + (\hat{\beta} - \beta_0)^\top D_{\beta\beta} \operatorname{ssr}^n(\mathbf{y}, \beta^*). \quad (5.27)$$

Ici  $\beta^*$  est une combinaison convexe de  $\hat{\beta}$  et  $\beta_0$ , qui peut être différente pour chaque ligne de l'équation, comme le demande le Théorème de Taylor.

Notre prochaine étape consiste à examiner la limite du membre de droite de (5.27) quand  $n \rightarrow \infty$ . La matrice Hessienne  $D_{\beta\beta} \operatorname{ssr}^n(\mathbf{y}, \beta)$ , évaluée en un vecteur  $\beta \in \Theta$  quelconque, peut s'écrire comme

$$D_{\beta\beta} \operatorname{ssr}^n(\mathbf{y}, \beta) = \frac{1}{n} \sum_{t=1}^n D_{\beta\beta}(y_t - x_t(\beta))^2 = \frac{1}{n} \sum_{t=1}^n \mathbf{H}_t(y_t, \beta). \quad (5.28)$$

Cette forme est compatible avec l'application d'une loi des grands nombres, d'où la condition (i) du Théorème 5.2. Nous pouvons aussi conclure que

$$\operatorname{plim}_0 \left( D_{\beta\beta} \operatorname{ssr}^n(\mathbf{y}, \beta) \right) = D_{\beta\beta} \overline{\operatorname{ssr}}(\beta, \mu_0). \quad (5.29)$$

Pour saisir cette conclusion, souvenons-nous que la condition WULLN permet de préserver l'intégrabilité lorsque l'on passe à la limite quand  $n \rightarrow \infty$ . La suite  $\{D_{\beta\beta} \operatorname{ssr}^n(\mathbf{y}, \beta)\}$  peut être intégrée deux fois et conduira au résultat  $\{\operatorname{ssr}^n(\mathbf{y}, \beta)\}$ , qui converge vers  $\overline{\operatorname{ssr}}(\beta, \mu_0)$  sous  $\mu_0$ . Ainsi, la limite de la double intégrale de  $\{D_{\beta\beta} \operatorname{ssr}^n(\mathbf{y}, \beta)\}$  sous  $\mu_0$  est  $\overline{\operatorname{ssr}}(\beta, \mu_0)$ , et puisque  $\overline{\operatorname{ssr}}(\beta, \mu_0)$  ne peut posséder qu'une seule matrice Hessienne, nous obtenons (5.29).

Puisque  $\beta^*$  est une combinaison convexe de  $\hat{\beta}$  et de  $\beta_0$ , et que  $\hat{\beta}$  converge vers  $\beta_0$ ,  $\beta^*$  doit aussi converger. Ainsi, compte tenu de l'uniformité de la convergence garantie par la condition WULLN,

$$\text{plim}_{n \rightarrow \infty} (D_{\beta\beta} \text{ssr}^n(\mathbf{y}, \beta^*)) = D_{\beta\beta} \overline{\text{ssr}}(\beta_0, \mu_0).$$

Si la condition (iii) du Théorème 5.2 est satisfaite, cette dernière matrice est définie positive et par conséquent est aussi non singulière et inversible. Cette condition peut être rattachée à la condition d'**identification asymptotique stricte**, puisqu'elle nécessite que (5.12) soit satisfaite mais également que la condition suffisante du second ordre soit satisfaite au minimum. Le résultat est que sous la condition (iii) nous pouvons récrire (5.27) comme

$$\hat{\beta} - \beta_0 = -(D_{\beta\beta} \text{ssr}^n(\mathbf{y}, \beta^*))^{-1} D_{\beta}^{\top} \text{ssr}^n(\mathbf{y}, \beta_0), \quad (5.30)$$

où la matrice inverse du membre de droite existe avec une probabilité arbitrairement proche de un, pour un  $n$  assez grand, et satisfait

$$\text{plim}_{n \rightarrow \infty} (D_{\beta\beta} \text{ssr}^n(\mathbf{y}, \beta^*))^{-1} = (D_{\beta\beta} \overline{\text{ssr}}(\beta_0, \mu_0))^{-1}. \quad (5.31)$$

L'argument du paragraphe précédent a utilisé une loi des grands nombres. Si nous multiplions (5.30) par  $n^{1/2}$ , nous pouvons aussi utiliser un théorème de la limite centrale. Le résultat de la multiplication est

$$n^{1/2}(\hat{\beta} - \beta_0) = -(D_{\beta\beta} \text{ssr}^n(\mathbf{y}, \beta^*))^{-1} (n^{1/2} D_{\beta}^{\top} \text{ssr}^n(\mathbf{y}, \beta_0)). \quad (5.32)$$

Le second facteur du membre de droite de cette équation est

$$\begin{aligned} n^{1/2} D_{\beta}^{\top} \text{ssr}^n(\mathbf{y}, \beta_0) &= n^{-1/2} \sum_{t=1}^n D_{\beta}^{\top} (y_t - x_t(\beta_0))^2 \\ &= -2n^{-1/2} \sum_{t=1}^n \mathbf{X}_t^{\top}(\beta_0) u_t. \end{aligned} \quad (5.33)$$

La raison de l'existence de la condition (ii) du Théorème 5.2 est maintenant claire: sous cette condition (5.33) a une **distribution asymptotique** normale, d'espérance nulle. De plus, sa matrice de covariance limite sera

$$\lim_{n \rightarrow \infty} \left( \frac{4}{n} \sum_{t=1}^n E_0(u_t^2 \mathbf{X}_t^{\top}(\beta_0) \mathbf{X}_t(\beta_0)) \right).$$

Mais puisque pour n'importe quel  $\beta$ ,  $u_t$  est indépendant de  $x_t(\beta)$  et par conséquent aussi de  $\mathbf{X}_t(\beta)$ , cette matrice de covariance devient simplement

$$4\sigma_0^2 \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}_0^{\top} \mathbf{X}_0). \quad (5.34)$$

La condition (i) de Théorème 5.2 nous permet de supposer que la condition WULLN est valable pour  $\{n^{-1} \sum_{t=1}^n \mathbf{H}_t(\mathbf{y}_t, \boldsymbol{\beta})\}$ . A partir de (5.24) nous voyons que  $\{n^{-1} \sum_{t=1}^n \mathbf{X}_t^\top(\boldsymbol{\beta}) \mathbf{X}_t(\boldsymbol{\beta})\}$  est en effet une composante de  $\{n^{-1} \sum_{t=1}^n \mathbf{H}_t(\mathbf{y}_t, \boldsymbol{\beta})\}$ , de sorte que nous pouvons supposer que la condition WULLN s'y applique également.

Nous en savons à présent suffisamment pour calculer la distribution limite du membre de gauche de (5.32), à savoir,  $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ . A partir de (5.31) et de (5.34) nous voyons que cette distribution est normale, d'espérance nulle et de matrice de covariance égale à

$$4\sigma_0^2 (D_{\beta\beta} \overline{ssr}(\boldsymbol{\beta}_0, \mu_0))^{-1} \text{plim}_0 (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0) (D_{\beta\beta} \overline{ssr}(\boldsymbol{\beta}_0, \mu_0))^{-1}. \quad (5.35)$$

Cette expression peut se simplifier. A partir de (5.24) et de (5.28),

$$\begin{aligned} D_{\beta\beta} \overline{ssr}^n(\mathbf{y}, \boldsymbol{\beta}_0) &= \frac{1}{n} \sum_{t=1}^n \mathbf{H}_t(\mathbf{y}_t, \boldsymbol{\beta}_0) \\ &= \frac{1}{n} \sum_{t=1}^n 2(\mathbf{X}_t^\top(\boldsymbol{\beta}_0) \mathbf{X}_t(\boldsymbol{\beta}_0) - \mathbf{A}_t(\boldsymbol{\beta}_0) u_t). \end{aligned} \quad (5.36)$$

Du fait que  $u_t$  et  $\mathbf{A}_t(\boldsymbol{\beta}_0)$  sont indépendants, tout comme  $u_t$  et  $\mathbf{X}_t(\boldsymbol{\beta}_0)$ ,

$$E_0(\mathbf{A}_t(\boldsymbol{\beta}_0) u_t) = \mathbf{0}. \quad (5.37)$$

Puisque la condition (i) du Théorème 5.2 nous permet l'usage d'une loi des grands nombres sur (5.36), il s'ensuit de (5.37) que

$$\begin{aligned} D_{\beta\beta} \overline{ssr}(\boldsymbol{\beta}_0, \mu_0) &= \text{plim}_0 (D_{\beta\beta} \overline{ssr}^n(\mathbf{y}, \boldsymbol{\beta}_0)) \\ &= \text{plim}_0 \left( \frac{1}{n} \sum_{t=1}^n 2(\mathbf{X}_t^\top(\boldsymbol{\beta}_0) \mathbf{X}_t(\boldsymbol{\beta}_0) - \mathbf{A}_t(\boldsymbol{\beta}_0) u_t) \right) \\ &= 2 \text{plim}_0 (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0). \end{aligned} \quad (5.38)$$

Par conséquent, la matrice de covariance limite de  $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$  est, à partir de (5.35) et de (5.38),

$$\sigma_0^2 \text{plim}_0 (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1}.$$

Puisque ceci est l'expression (5.25), nous avons démontré la dernière partie du Théorème 5.2.

Il sera utile d'exprimer (5.32) au vu de (5.33) et de (5.38). Elle devient

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u} + o(1). \quad (5.39)$$



Dans le cas d'un modèle de régression linéaire avec  $\mathbf{x}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ , l'égalité serait exacte sans le terme  $o(1)$ . Tous les facteurs des puissances de  $n$  sont inutiles dans ce cas, et nous obtenons le résultat familier

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}. \quad (5.40)$$

Le résultat (5.39) s'interprète donc comme la contrepartie asymptotique de (5.40) pour les modèles de régression non linéaires. Nous utiliserons (5.39) ainsi que d'autres résultats de cette section dans la Section 5.6 afin d'établir les propriétés des résidus NLS. Avant cela, dans la prochaine section, nous étudierons une autre propriété importante de l'estimateur NLS, l'**efficacité asymptotique**.

## 5.5 EFFICACITÉ ASYMPTOTIQUE DES NLS

Jusqu'ici, nous n'avons rien dit sur les avantages de l'estimateur OLS ou NLS par rapport aux autres estimateurs. Un estimateur est dit plus **efficace** qu'un autre si, en moyenne, il fournit des estimations plus précises que l'autre. La raison de cette terminologie est qu'un estimateur qui fournit des estimations plus précises utilise l'information disponible dans l'échantillon de façon plus efficace. Nous pourrions définir l'efficacité d'autant de manières différentes que nous pourrions imaginer évaluer la précision relative de deux estimateurs, et il existe ainsi de nombreuses définitions de l'efficacité dans la littérature. Nous ne traiterons ici que des deux définitions les plus largement utilisées.

Supposons que  $\hat{\boldsymbol{\theta}}$  et  $\check{\boldsymbol{\theta}}$  soient deux estimateurs sans biais d'un vecteur de dimension  $k$  des paramètres  $\boldsymbol{\theta}$ , que la véritable valeur soit  $\boldsymbol{\theta}_0$ , et que les deux estimateurs aient des matrices de covariance

$$\begin{aligned} \mathbf{V}(\hat{\boldsymbol{\theta}}) &\equiv E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top \quad \text{et} \\ \mathbf{V}(\check{\boldsymbol{\theta}}) &\equiv E(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top, \end{aligned}$$

respectivement. Alors nous avons:

*Définition 5.5.*

L'estimateur sans biais  $\hat{\boldsymbol{\theta}}$ , de matrice de covariance  $\mathbf{V}(\hat{\boldsymbol{\theta}})$ , est dit plus efficace que l'estimateur sans biais  $\check{\boldsymbol{\theta}}$ , de matrice de covariance  $\mathbf{V}(\check{\boldsymbol{\theta}})$ , si et seulement si  $\mathbf{V}(\check{\boldsymbol{\theta}}) - \mathbf{V}(\hat{\boldsymbol{\theta}})$ , la différence des deux matrices de covariance, est une matrice semi-définie positive.

Si  $\hat{\boldsymbol{\theta}}$  est plus efficace que  $\check{\boldsymbol{\theta}}$  au sens de cette définition, alors chaque composante du vecteur  $\boldsymbol{\theta}$ , et chaque combinaison linéaire de ces composantes, est estimé au moins aussi efficacement par  $\hat{\boldsymbol{\theta}}$  que par  $\check{\boldsymbol{\theta}}$ , ce qui signifie que la variance de l'estimateur basé sur  $\hat{\boldsymbol{\theta}}$  n'est jamais supérieur à celle de l'estimateur basé sur  $\check{\boldsymbol{\theta}}$ . Pour comprendre cela, considérons une combinaison linéaire quelconque des paramètres de  $\boldsymbol{\theta}$ , disons  $\mathbf{w}^\top \boldsymbol{\theta}$ , où  $\mathbf{w}$  est un vecteur de dimension  $k$ . Alors, les

variances des deux estimations de cette combinaison linéaire sont  $\mathbf{w}^\top \mathbf{V}(\check{\boldsymbol{\theta}})\mathbf{w}$  et  $\mathbf{w}^\top \mathbf{V}(\hat{\boldsymbol{\theta}})\mathbf{w}$ , de sorte que la différence entre elles est

$$\mathbf{w}^\top \mathbf{V}(\check{\boldsymbol{\theta}})\mathbf{w} - \mathbf{w}^\top \mathbf{V}(\hat{\boldsymbol{\theta}})\mathbf{w} = \mathbf{w}^\top (\mathbf{V}(\check{\boldsymbol{\theta}}) - \mathbf{V}(\hat{\boldsymbol{\theta}}))\mathbf{w}.$$

Puisque  $\mathbf{V}(\check{\boldsymbol{\theta}}) - \mathbf{V}(\hat{\boldsymbol{\theta}})$  est une matrice semi-définie positive, cette grandeur doit être soit positive soit nulle. Ainsi, quel que soit le paramètre ou la combinaison linéaire des paramètres que nous tentons d'estimer, nous pouvons être sûrs que  $\hat{\boldsymbol{\theta}}$  fournira un estimateur au moins aussi bon que  $\check{\boldsymbol{\theta}}$  si la différence entre leurs matrices de covariance est semi-définie positive. Dans la pratique, quand un estimateur est plus efficace qu'un autre, cette différence de matrices est très souvent définie positive. Lorsque c'est le cas, *chaque* paramètre ou combinaison linéaire des paramètres sera en fait estimé plus efficacement en utilisant  $\hat{\boldsymbol{\theta}}$ .

Quand nous estimons des modèles de régression non linéaire et d'autres types de modèles non linéaires, nous rencontrons rarement des estimations sans biais, et sommes rarement capables d'évaluer les matrices de covariance en échantillon fini des estimateurs. Il est par conséquent naturel de chercher un concept asymptotique comparable à l'efficacité dans le cas d'un échantillon fini. Le concept approprié est celui de l'**efficacité asymptotique**, définie comme suit:

*Définition 5.6.*

Supposons que  $\hat{\boldsymbol{\theta}}$  et  $\check{\boldsymbol{\theta}}$  soient deux estimateurs convergents du même vecteur paramétrique  $\boldsymbol{\theta}$ . Soient

$$\mathbf{V}^\infty(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) \equiv \lim_{n \rightarrow \infty} E_0(n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top) \quad \text{et}$$

$$\mathbf{V}^\infty(n^{1/2}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) \equiv \lim_{n \rightarrow \infty} E_0(n(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top)$$

les matrices de covariances asymptotiques de ces deux estimateurs. Alors l'estimateur  $\hat{\boldsymbol{\theta}}$  est asymptotiquement plus efficace que l'estimateur  $\check{\boldsymbol{\theta}}$  si

$$\mathbf{V}^\infty(n^{1/2}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) - \mathbf{V}^\infty(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0))$$

est une matrice semi-définie positive.

Un résultat célèbre sur l'efficacité est le **Théorème de Gauss-Markov**. Ce théorème s'applique au modèle de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}, \quad (5.41)$$

où les régresseurs  $\mathbf{X}$  sont fixes ou peuvent être traités comme fixes parce que nous conditionnons l'espérance de la variable dépendante par rapport à eux (consulter la Section 3.5). Ce théorème enseigne que:

*Théorème 5.3. Théorème de Gauss-Markov.*

L'estimateur OLS  $\hat{\beta} \equiv (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  des paramètres  $\beta$  du modèle de régression linéaire (5.41) est le **meilleur estimateur linéaire sans biais**, ou **BLUE**. Ceci signifie que si  $\mathbf{V}(\hat{\beta})$  est la matrice de covariance de  $\hat{\beta}$  sous un DGP appartenant au modèle (5.41), et si  $\mathbf{V}(\check{\beta})$  la matrice de covariance d'un autre estimateur quelconque sans biais  $\check{\beta}$ , dépendant linéairement du vecteur  $\mathbf{y}$ , alors  $\mathbf{V}(\check{\beta}) - \mathbf{V}(\hat{\beta})$  est une matrice semi-définie positive.

La démonstration de ce théorème est à la fois simple et instructive. Puisque  $\check{\beta}$  est une fonction linéaire de  $\mathbf{y}$ , nous pouvons l'écrire comme

$$\check{\beta} = \mathbf{A}\mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} + \mathbf{C}\mathbf{y}, \quad (5.42)$$

où  $\mathbf{C}$  est définie comme  $\mathbf{A} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ . Nous supposons que les données sont générées par un DGP qui est un cas particulier de (5.41), avec  $\beta = \beta_0$  et  $\sigma^2 = \sigma_0^2$ . Nous pouvons ainsi substituer  $\mathbf{X}\beta_0 + \mathbf{u}$  à  $\mathbf{y}$  dans (5.42) pour obtenir

$$\begin{aligned} \check{\beta} &= ((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top + \mathbf{C})(\mathbf{X}\beta_0 + \mathbf{u}) \\ &= \beta_0 + \mathbf{C}\mathbf{X}\beta_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} + \mathbf{C}\mathbf{u}. \end{aligned} \quad (5.43)$$

Il est clair à partir de (5.43) que  $E(\check{\beta})$  ne peut être égale à  $\beta_0$  que si  $\mathbf{C}\mathbf{X}\beta_0$  est égal au vecteur nul. Ceci peut être garanti pour toutes les valeurs de  $\beta_0$  à la condition que  $\mathbf{C}\mathbf{X} = \mathbf{0}$ . Ainsi, la contrainte que  $\check{\beta}$  soit un estimateur linéaire *sans biais* implique que, tout d'abord, le second terme du membre de droite de (5.42),  $\mathbf{C}\mathbf{y}$ , ait une espérance nulle parce que  $\mathbf{C}\mathbf{X}\beta_0 = \mathbf{0}$ , et ensuite, que les deux termes du membre de droite de (5.42) aient une covariance nulle. Pour voir ce second point, observons que

$$\begin{aligned} E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \mathbf{y}^\top \mathbf{C}^\top) &= E\left((\beta_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}) \mathbf{u}^\top \mathbf{C}^\top\right) \\ &= \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{C}^\top \\ &= \mathbf{0}. \end{aligned} \quad (5.44)$$

Par conséquent, l'équation (5.42) indique que l'estimateur linéaire sans biais  $\check{\beta}$  est égal à l'estimateur des moindres carrés  $\hat{\beta}$  auquel s'ajoute un élément aléatoire  $\mathbf{C}\mathbf{y}$  qui est non corrélé à  $\hat{\beta}$ . Comme nous le verrons plus tard et dans le Chapitre 8, c'est un phénomène que l'on observe assez généralement: asymptotiquement, un estimateur non efficace est toujours égal à un estimateur efficace auquel s'ajoute un bruit aléatoire qui lui est indépendant.

Le résultat (5.44) démontre en grande partie le Théorème de Gauss-Markov, puisqu'il implique que

$$\begin{aligned} &E(\check{\beta} - \beta_0)(\check{\beta} - \beta_0)^\top \\ &= E\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} + \mathbf{C}\mathbf{u}\right)\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} + \mathbf{C}\mathbf{u}\right)^\top \\ &= \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1} + \sigma_0^2 \mathbf{C}\mathbf{C}^\top. \end{aligned} \quad (5.45)$$

Ainsi, la différence entre les matrices de covariance  $\check{\beta}$  et  $\hat{\beta}$  est  $\sigma_0^2 \mathbf{C} \mathbf{C}^\top$ , qui est une matrice semi-définie positive. Notons que l'hypothèse que  $E(\mathbf{u} \mathbf{u}^\top) = \sigma_0^2 \mathbf{I}$  est ici cruciale. Si à la place nous avions  $E(\mathbf{u} \mathbf{u}^\top) = \mathbf{\Omega}$ , avec une matrice  $\mathbf{\Omega}$  définie positive quelconque de dimension  $n \times n$ , la dernière ligne de (5.45) serait

$$\begin{aligned} & (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Omega} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ & + \mathbf{C} \mathbf{\Omega} \mathbf{C}^\top + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Omega} \mathbf{C}^\top + \mathbf{C} \mathbf{\Omega} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}, \end{aligned}$$

et nous ne pourrions rien conclure sur l'efficacité relative de  $\check{\beta}$  et de  $\hat{\beta}$ .

Comme un exemple simple d'application du Théorème de Gauss-Markov, supposons que  $\check{\beta}$  soit l'estimateur OLS obtenu en régressant  $\mathbf{y}$  sur  $\mathbf{X}$  et  $\mathbf{Z}$  conjointement, où  $\mathbf{Z}$  est une matrice de régresseurs telle que  $E(\mathbf{y} | \mathbf{X}, \mathbf{Z}) = E(\mathbf{y} | \mathbf{X}) = \mathbf{X} \boldsymbol{\beta}$ . Puisque l'information que  $\mathbf{Z}$  n'appartient pas à la régression est ignorée quand nous construisons  $\check{\beta}$ , ce dernier doit être en général non efficace. En utilisant le Théorème FWL, nous trouvons que

$$\check{\beta} = (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_Z \mathbf{y}, \quad (5.46)$$

où, comme d'habitude,  $\mathbf{M}_Z$  est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{Z})$ . Si nous écrivons  $\check{\beta}$  sous la forme (5.42), nous obtenons

$$\begin{aligned} (5.47) \quad \check{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} + ((\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_Z - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{M}_Z - \mathbf{X}^\top \mathbf{M}_Z \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \left( \mathbf{X}^\top \mathbf{M}_Z (\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \right) \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_Z \mathbf{M}_X \mathbf{y} \\ &= \hat{\beta} + \mathbf{C} \mathbf{y}. \end{aligned}$$

Ainsi, dans ce cas, la matrice  $\mathbf{C}$  est  $(\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_Z \mathbf{M}_X$ . Nous voyons que l'estimateur inefficace  $\check{\beta}$  est égal à l'estimateur efficace  $\hat{\beta}$  plus un élément aléatoire qui lui est non corrélé. L'absence de corrélation entre  $\hat{\beta}$  et  $\mathbf{C} \mathbf{y}$  provient du fait que  $\mathbf{C} \mathbf{X} = \mathbf{0}$  (nécessaire pour que  $\mathbf{C} \mathbf{y}$  ait une espérance nulle), qui est vraie parce que  $\mathbf{M}_X$  annule  $\mathbf{X}$ . De plus, nous voyons que

$$\begin{aligned} E(\check{\beta} - \beta_0)(\check{\beta} - \beta_0)^\top &= \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1} \\ &+ \sigma_0^2 (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_Z \mathbf{M}_X \mathbf{M}_Z \mathbf{X} (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1}. \end{aligned} \quad (5.48)$$

Le second terme du membre de droite de (5.48) sera en général une matrice semi-définie positive, comme prévu. Ce sera une matrice nulle si  $\mathbf{Z}$  est orthogonale à  $\mathbf{X}$ , au quel cas  $\mathbf{M}_Z \mathbf{X} = \mathbf{X}$  et  $\mathbf{M}_X \mathbf{M}_Z \mathbf{X} = \mathbf{0}$ . Ainsi, nous obtenons le résultat familier qui est que l'addition des variables explicatives qui n'appartiennent pas à une régression réduira l'efficacité sauf dans le cas

rare où les variables supplémentaires sont orthogonales à celles qui appartiennent au modèle.

Il est important de garder à l'esprit les limites du Théorème de Gauss-Markov. Il *n'enseigne pas* que l'estimateur OLS  $\hat{\beta}$  est meilleur que tout autre estimateur concevable. Des estimateurs non linéaires et/ou biaisés peuvent être plus performants que l'estimateur OLS dans certaines circonstances. En particulier, comme nous le verrons dans le Chapitre 8, seule l'hypothèse de normalité des aléas fera en général coïncider l'estimateur OLS avec l'estimateur du maximum de vraisemblance, qui sera asymptotiquement “meilleur” sous des conditions assez générales lorsque la distribution des aléas est connue. De plus, le théorème ne s'applique qu'à un modèle correctement spécifié avec des erreurs homoscedastiques.

Pour comprendre l'importance d'une spécification correcte, reconsidérons l'exemple de la régression linéaire dans laquelle  $E(\mathbf{y} | \mathbf{X}, \mathbf{Z}) = \mathbf{X}\beta$ . Si nous ne savons pas que l'espérance de  $\mathbf{y}$  conditionnellement à  $\mathbf{X}$  et  $\mathbf{Z}$  est indépendante de  $\mathbf{Z}$ , il est raisonnable d'estimer le modèle de régression

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{u}. \quad (5.49)$$

L'estimateur OLS de  $\beta$ , (5.46), est, d'après le Théorème de Gauss-Markov, asymptotiquement efficace pour le modèle complet (5.49), qui admet des DGP pour lesquels  $\gamma$  est non nul. Mais il est inefficace relativement à l'estimateur  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  pour la classe des DGP où  $\gamma = \mathbf{0}$ . Cependant, ceci constitue une classe *restreinte* des DGP, et l'estimateur  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  est en général *non convergent* pour des DGP satisfaisant (5.49) avec  $\gamma \neq \mathbf{0}$ . Son efficacité supérieure a été obtenue en supposant que  $\gamma = \mathbf{0}$ , et en risquant la non convergence si cette hypothèse est fausse.

Il est évident que le Théorème de Gauss-Markov ne peut pas s'appliquer à l'estimateur NLS, puisque cet estimateur n'est en général ni linéaire ni sans biais. Néanmoins, il est asymptotiquement efficace (Définition 5.6) dans un certain sens. Souvenons-nous du résultat (5.39) de la Section 5.4, que nous pouvons récrire comme

$$n^{1/2}(\hat{\beta} - \beta_0) \stackrel{a}{=} \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u}.$$

Il est possible de considérer une classe d'estimateurs, que nous pouvons à nouveau noter  $\check{\beta}$ , possédant la propriété que

$$n^{1/2}(\check{\beta} - \beta_0) \stackrel{a}{=} \left( \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} (n^{-1/2} \mathbf{X}_0^\top) + n^{-1/2} \mathbf{C} \right) \mathbf{u}, \quad (5.50)$$

où chaque élément de la matrice  $\mathbf{C}$  de dimension  $k \times n$  (qui peut dépendre de  $\beta$ ) est  $O(1)$ , et où nous supposons que

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{C} \mathbf{u}) = \mathbf{0} \quad \text{et} \quad \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{C} \mathbf{X}_0) = \mathbf{0}.$$

Notons que, conformément à (5.50),  $n^{1/2}(\check{\beta} - \beta_0)$  est asymptotiquement une fonction linéaire du vecteur des aléas  $\mathbf{u}$ . Ces hypothèses sont suffisantes pour garantir que  $\check{\beta}$  est convergent si  $\hat{\beta}$  l'est. La démonstration que

$$\mathbf{V}^\infty(n^{1/2}(\check{\beta} - \beta_0)) - \mathbf{V}^\infty(n^{1/2}(\hat{\beta} - \beta_0))$$

est une matrice semi-définie positive est alors un exercice très comparable à la démonstration du Théorème de Gauss-Markov. Par conséquent, nous concluons que l'estimateur NLS est asymptotiquement plus efficace que n'importe quel autre estimateur de la forme (5.50). Nous nous référerons à de tels estimateurs en tant qu'estimateurs **convergents et asymptotiquement linéaires**, l'estimateur NLS étant considéré comme le meilleur estimateur convergent et asymptotiquement linéaire.

Ce résultat peut ne pas sembler très significatif parce que, jusqu'à présent, nous n'avons pas vu d'autres estimateurs convergents et asymptotiquement linéaires. Cependant, il devrait être clair à partir de la similitude de l'estimateur NLS et OLS que si nous estimions le modèle

$$\mathbf{y} = \mathbf{x}(\beta, \gamma) + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I},$$

où  $\mathbf{x}(\beta, \mathbf{0}) = \mathbf{x}(\beta)$ , nous obtiendrions un estimateur qui satisfait (5.50) asymptotiquement. La forme particulière de (5.50) serait similaire à l'expression (5.47) dans le cas linéaire. Parmi les estimateurs convergents et asymptotiquement linéaires, nous trouvons l'estimateur des moindres carrés généralisés non linéaires que nous examinerons dans le Chapitre 9 et l'estimateur des variables instrumentales non linéaire qui sera examiné dans le Chapitre 7.

Un résultat plus fort sur l'efficacité des NLS est disponible si nous supposons que les aléas sont normalement distribués. Dans ce cas, l'estimateur NLS du vecteur de paramètres  $\beta$  correspond à l'estimateur du maximum de vraisemblance. Comme nous le verrons dans le Chapitre 8, l'estimateur ML est asymptotiquement efficace dans un sens très fort, pourvu que toute la structure stochastique du modèle soit correctement spécifiée. Cela implique que l'estimateur NLS est asymptotiquement efficace par rapport à une classe très large de techniques d'estimation pour la classe des modèles de régression non linéaire avec des perturbations homoscédastiques, indépendantes et normalement distribuées.

## 5.6 PROPRIÉTÉS DES RÉSIDUS NLS

Nous avons jusqu'à maintenant discuté de la plupart des points intéressants qui concernent les propriétés asymptotiques de l'estimateur des moindres carrés non linéaires. Dans cette section, nous souhaitons discuter des propriétés des **résidus NLS**, à savoir la suite  $\{y_t - \hat{x}_t\}$ . Ces propriétés sont importantes pour plusieurs raisons, et non pas seulement parce que les résidus seront utilisés pour estimer la variance des aléas  $\sigma^2$ .

Afin d'obtenir les propriétés asymptotiques des résidus NLS, nous commençons par opérer un développement de Taylor sur un résidu type autour de  $\beta = \beta_0$ . Ce développement est

$$\begin{aligned}\hat{u}_t &\equiv y_t - x_t(\hat{\beta}) = y_t - x_{0t} - \mathbf{X}_t^*(\hat{\beta} - \beta_0) \\ &= u_t - \mathbf{X}_t^*(\hat{\beta} - \beta_0),\end{aligned}\tag{5.51}$$

où, comme d'habitude,  $\mathbf{X}_t^* \equiv \mathbf{X}_t(\beta^*)$  pour une combinaison convexe quelconque  $\beta^*$  de  $\hat{\beta}$  et de  $\beta_0$ . Sous les conditions du Théorème 5.2,  $\hat{\beta} - \beta_0 = O(n^{-1/2})$ . Ainsi, nous pouvons conclure immédiatement que

$$\hat{u}_t = u_t + O(n^{-1/2}),\tag{5.52}$$

qui implique que les résidus estiment de façon convergente les vrais aléas.

Le résultat simple (5.52) est extrêmement précieux, mais il n'est pas suffisamment détaillé pour tous les cas de figures. Pour s'en convaincre, considérons l'expression

$$n^{-1/2} \mathbf{a}^\top \hat{\mathbf{u}} = n^{-1/2} \sum_{t=1}^n a_t \hat{u}_t\tag{5.53}$$

pour un vecteur  $\mathbf{a}$  quelconque dont les éléments forment une suite non stochastique  $\{a_t\}$ . Si chaque  $a_t$  est de l'ordre de l'unité, alors la substitution de (5.52) dans (5.53) montre que cette dernière est égale à

$$n^{-1/2} \sum_{t=1}^n a_t u_t + n^{-1/2} \sum_{t=1}^n O(n^{-1/2}).\tag{5.54}$$

Si un théorème de la limite centrale s'applique au premier terme de (5.54), alors ce terme est de l'ordre de l'unité. Mais le second terme l'est aussi, compte tenu de la somme de  $n$  termes, et ne peut donc pas être ignoré si nous souhaitons établir les propriétés de (5.53). Ceci constitue un résultat extrêmement important, du fait que nous nous intéresserons très souvent dans l'analyse asymptotique à des grandeurs comparables à (5.53). Le résultat nous enseigne que pour de tels cas nous ne pouvons pas ignorer la distinction entre les aléas  $u_t$  et les résidus  $\hat{u}_t$ .

Pour cette raison, nous allons affiner le résultat (5.52). Pour y parvenir, nous utilisons le résultat fondamental de la normalité asymptotique (5.39). Il peut se récrire comme

$$\hat{\beta} - \beta_0 = n^{-1/2} \left( (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u} + o(1) \right).\tag{5.55}$$

La substitution de (5.55) dans la seconde ligne de (5.51) conduit à

$$\hat{u}_t = u_t - n^{-1/2} \mathbf{X}_t^* (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u} + o(n^{-1/2}).\tag{5.56}$$

Il devrait être clair que le premier terme est  $O(1)$  et que le second est  $O(n^{-1/2})$ . Ainsi, (5.56) livre les deux premiers termes dans ce que nous appelons le **développement stochastique** du résidu  $\hat{u}_t$ . Mais ce développement est encore inutilement compliqué, parce que nous avons

$$\mathbf{X}_t^* = \mathbf{X}_{0t} + \mathbf{A}_t^*(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{X}_{0t} + O(n^{-1/2})$$

grâce au Théorème de Taylor et au fait que  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O(n^{-1/2})$ ; souvenons-nous que  $\mathbf{A}_t$  est la matrice Hessienne de la fonction de régression  $x_t(\boldsymbol{\beta})$ . Ainsi (5.56) peut s'écrire plus simplement comme

$$\hat{u}_t = u_t - n^{-1/2} \mathbf{X}_{0t} (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u} + o(n^{-1/2}).$$

Puisque ceci est vrai pour tout  $t$ , nous avons l'équation vectorielle

$$\hat{\mathbf{u}} = \mathbf{u} - \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{u} + o(n^{-1/2}),$$

où le dernier terme doit à présent s'interpréter comme un vecteur de dimension  $n$ , dont chaque composante est  $o(n^{-1/2})$ . Cette équation peut s'écrire en terme de la projection  $\mathbf{P}_0 \equiv \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top$  et de sa projection complémentaire  $\mathbf{M}_0 \equiv \mathbf{I} - \mathbf{P}_0$ :

$$\hat{\mathbf{u}} = \mathbf{u} - \mathbf{P}_0 \mathbf{u} + o(n^{-1/2}) = \mathbf{M}_0 \mathbf{u} + o(n^{-1/2}). \quad (5.57)$$

Il s'agit de l'équivalent asymptotique du résultat exact indiquant que, pour des modèles linéaires, les résidus OLS sont la projection orthogonale des perturbations sur l'espace complémentaire de l'espace engendré par les régresseurs. Rappelons que si nous exécutons la régression  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , le DGP étant en fait un cas particulier de ce modèle, alors nous avons exactement

$$\hat{\mathbf{u}} = \mathbf{M}_X \mathbf{u}. \quad (5.58)$$

Le résultat (5.57) se résume à (5.58) quand le modèle est linéaire. La matrice de projection  $\mathbf{M}_0$  est alors égale à  $\mathbf{M}_X$ , et le terme  $o(n^{-1/2})$ , qui provenait uniquement de la non linéarité de  $\mathbf{x}(\boldsymbol{\beta})$ , disparaît.

Substituons à présent l'expression la plus à droite de (5.57) dans (5.53). Cette dernière devient

$$n^{-1/2} \mathbf{a}^\top \hat{\mathbf{u}} = n^{-1/2} \mathbf{a}^\top \mathbf{M}_0 \mathbf{u} + n^{-1/2} \sum_{t=1}^n o(n^{-1/2}). \quad (5.59)$$

Le premier terme du membre de droite est ici  $O(1)$ , tandis que le second est  $o(1)$ . Ainsi, contrairement à ce qui survenait quand nous remplacions  $\hat{u}_t$  par  $u_t$ , nous pouvons ignorer le second terme du membre de droite de (5.59).



Ainsi, le résultat (5.57) fournit ce dont nous avons besoin lorsque nous traitons de l'analyse asymptotique d'expressions comparables à (5.53).

Nous devrions nous attarder ici pour éclaircir la relation entre le résultat asymptotique (5.57), le résultat linéaire exact (5.58), et deux autres résultats. Ces autres résultats sont (1.03), qui établit que les résidus OLS sont orthogonaux aux régresseurs, et (2.05), que nous pouvons exprimer comme  $\hat{\mathbf{X}}^\top \hat{\mathbf{u}} = \mathbf{0}$ , et qui établit que les résidus NLS sont orthogonaux à  $\mathbf{X}(\hat{\boldsymbol{\beta}})$ . Cette seconde paire de résultats conduit aux propriétés *numériques* des OLS et des NLS qui doivent être vérifiées quel que soit le processus ayant généré les données. Par contraste, (5.57) et (5.58) sont des résultats *statistiques* qui ne sont valables que si le DGP appartient véritablement au modèle de régression adéquat. Aussi bien les OLS que les NLS opèrent ce que l'on pourrait appeler une projection orthogonale, mécanique et parfaite; c'est précisément ce qu'indiquent les résultats (1.03) et (2.05). Si de plus le DGP appartient au modèle linéaire ou non linéaire considéré, cette projection correspond à la projection du véritable vecteur d'aléas  $\mathbf{u}$  sur le complément orthogonal du sous-espace  $\mathcal{S}(\mathbf{X}_0)$ . C'est exactement parce que cette projection annule un nombre fixe et fini de directions ( $k$  dans la notation utilisée) que nous obtenons le résultat simple (5.52).

Les directions annulées de  $\mathbf{u}$ ,  $\mathbf{P}_0 \mathbf{u}$ , correspondent asymptotiquement (dans le cas linéaire, exactement) aux erreurs commises en estimant les paramètres. Pour comprendre ceci, nous pouvons récrire (5.55) comme

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \stackrel{a}{=} (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{u}, \quad (5.60)$$

d'où

$$\mathbf{X}_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} \mathbf{P}_0 \mathbf{u}. \quad (5.61)$$

En exprimant (5.57) à l'aide d'une notation simplifiée comparable, nous voyons que

$$\hat{\mathbf{u}} \stackrel{a}{=} \mathbf{u} - \mathbf{P}_0 \mathbf{u} = \mathbf{M}_0 \mathbf{u}. \quad (5.62)$$

Il est ainsi assez intuitif de conclure que, puisque la variance de l'estimateur  $\hat{\boldsymbol{\beta}}$  tend vers zéro quand la taille d'échantillon tend vers l'infini, les perturbations sont de mieux en mieux estimées par les résidus quand la taille d'échantillon augmente.

Les résultats asymptotiques (5.60), (5.61), et (5.62) sont les résultats *statistiques* essentiels sur lesquels se base l'étude asymptotique des NLS, si le DGP est supposé appartenir au modèle de régression non linéaire estimé. Naturellement, si ce n'est pas le cas, ces résultats ne sont plus valables. Nous ferons donc cette hypothèse la plupart du temps, bien qu'elle soit non plausible. Cependant, quand nous discuterons des sources de la puissance des tests dans le Chapitre 12, nous examinerons ce qui survient quand le DGP appartient *quasiment* au modèle estimé. Et dans le Chapitre 11, dans le contexte de ce que l'on appelle tests d'hypothèses non emboîtées, nous rencontrerons

un cas où l'analyse dépend des propriétés des valeurs ajustées des moindres carrés quand le DGP n'appartient pas du tout au modèle estimé.

Une utilisation importante des résidus  $\hat{u}_t$  consiste à estimer la variance des erreurs  $\sigma^2$ . Les deux estimateurs principaux suggérés dans le Chapitre 2 étaient

$$\hat{\sigma}^2 \equiv \frac{1}{n} \sum_{t=1}^n (y_t - x_t(\hat{\beta}))^2 \quad \text{et}$$

$$s^2 \equiv \frac{1}{n-k} \sum_{t=1}^n (y_t - x_t(\hat{\beta}))^2.$$

Nous démontrerons que ces deux estimateurs convergent mais que  $s^2$  est préférable à  $\hat{\sigma}^2$ .

Le résultat asymptotique fondamental pour les résidus NLS, l'équation (5.57), peut être récrit comme

$$\hat{\mathbf{u}} = \mathbf{M}_0 \mathbf{u} + o(n^{-1/2}) \mathbf{a} \quad (5.63)$$

pour un vecteur aléatoire  $\mathbf{a}$  quelconque de dimension  $n$ , dont chaque élément est  $O(1)$ . Ici, la notation signifie que chaque élément de  $\mathbf{a}$  est multiplié par un scalaire qui est  $o(n^{-1/2})$ . En utilisant (5.63), nous voyons que

$$\begin{aligned} \hat{\sigma}^2 &\equiv n^{-1} \hat{\mathbf{u}}^\top \hat{\mathbf{u}} \\ &= n^{-1} \mathbf{u}^\top \mathbf{M}_0 \mathbf{u} + 2n^{-1} o(n^{-1/2}) \mathbf{a}^\top \mathbf{M}_0 \mathbf{u} + n^{-1} o(n^{-1}) \mathbf{a}^\top \mathbf{a} \\ &= n^{-1} \mathbf{u}^\top \mathbf{u} - n^{-1} \mathbf{u}^\top \mathbf{P}_0 \mathbf{u} + 2o(n^{-3/2}) \mathbf{a}^\top \mathbf{M}_0 \mathbf{u} + o(n^{-2}) \mathbf{a}^\top \mathbf{a}. \end{aligned} \quad (5.64)$$

La dernière ligne peut conduire à plusieurs résultats intéressants.

Le premier terme dans la dernière ligne de (5.64) est à l'évidence  $O(1)$ . De plus, puisque ce terme est juste  $n^{-1}$  fois la somme de  $n$  aléas au carré indépendants, une loi des grands nombres peut s'y appliquer sous des conditions de régularités amoindries. Ainsi, la limite en probabilité de ce premier terme est simplement  $\sigma_0^2$ . Nous montrerons dans un moment que les trois autres termes dans la dernière ligne de (5.64) sont soit  $O(n^{-1})$  soit  $o(n^{-1})$ . Ainsi, la limite en probabilité de  $\hat{\sigma}^2$  est simplement la limite en probabilité du premier terme, de sorte que nous concluons que  $\hat{\sigma}^2$  converge vers  $\sigma_0^2$ . Comme  $s^2 = (n/(n-k))\hat{\sigma}^2$  et que la limite en probabilité de  $n/(n-k)$  est l'unité,  $s^2$  est évidemment un estimateur convergent.

Le second terme dans la dernière ligne de (5.64) peut être récrit comme

$$n^{-1} (n^{-1/2} \mathbf{u}^\top \mathbf{X}_0) (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} (n^{-1/2} \mathbf{X}_0^\top \mathbf{u}).$$

C'est  $n^{-1}$  fois le produit de trois facteurs, chacun étant  $O(1)$ , ce qui implique que le second terme dans son ensemble doit être  $O(n^{-1})$ .

Les troisième et quatrième termes sont d'un ordre encore inférieur. Le quatrième terme est facile à traiter. La quantité  $\mathbf{a}^\top \mathbf{a}$  doit être  $O(n)$ , parce qu'elle est simplement la somme de  $n$  carrés, chacun étant  $O(1)$ . Ainsi, le quatrième terme est évidemment  $o(n^{-1})$ . Si les composantes du vecteur  $\mathbf{a}$  étaient non stochastiques, le traitement du troisième terme serait aussi facile. A l'aide d'arguments comparables à ceux utilisés en connexion avec (5.59), nous pourrions montrer que  $2n^{-1}\mathbf{a}^\top \mathbf{M}_0 \mathbf{u}$  est  $O(n^{-1/2})$ . Il s'ensuivrait alors immédiatement que le troisième terme serait  $o(n^{-1})$ . Le problème est que le vecteur  $\mathbf{a}$  n'est pas véritablement non stochastique, puisqu'il dépend de  $\hat{\beta}$ . Il est néanmoins possible de prouver que le troisième terme dans la dernière ligne de (5.64) est en fait d'un ordre inférieur à  $n^{-1}$ . La démonstration nécessite un développement de Taylor au second ordre de  $\hat{u}_t \equiv y_t - x_t(\hat{\beta})$  autour de  $\beta_0$  et utilise le résultat fondamental (5.39). Nous laissons cette démonstration en exercice.

En utilisant les résultats précédents sur les ordres des quatre termes dans la dernière ligne de (5.64), nous pouvons à présent comparer les propriétés de  $\hat{\sigma}^2$  et de  $s^2$  à l'ordre  $n^{-1}$ . Nous avons déjà vu que ces deux estimateurs sont convergents. En utilisant des techniques usuelles, similaires à celles utilisées dans (3.08), il est facile de montrer que  $E(n^{-1}\mathbf{u}^\top \mathbf{P}_0 \mathbf{u}) = (k/n)\sigma_0^2$ . Par conséquent, à l'ordre  $n^{-1}$ ,

$$E(\hat{\sigma}^2) = \frac{n-k}{n} \sigma_0^2.$$

Ainsi, comme nous le savions déjà,  $\hat{\sigma}^2$  est biaisée vers le bas. Par contraste, il est facile de voir que, pour le même ordre,  $s^2$  est sans biais. Ce résultat favorise fortement l'usage de  $s^2$  plutôt que de  $\hat{\sigma}^2$  lorsque nous estimons la variance des aléas d'un modèle de régression non linéaire, tout comme c'est le cas pour un modèle linéaire. Naturellement, le fait que  $s^2$  soit sans biais à l'ordre  $n^{-1}$  n'implique pas qu'il sera sans biais quel que soit l'ordre. En général, il sera biaisé pour un ordre inférieur à  $n^{-1}$ .

La démonstration de la convergence de  $s^2$  (ou de  $\hat{\sigma}^2$ ) était très différente de celle de  $\hat{\beta}$ , parce que bien que  $\sigma^2$  soit un paramètre du modèle de régression non linéaire, ce n'est pas un argument de la fonction somme-des-carrés. Comme nous l'avons mentionné plus tôt, le paramètre  $\sigma^2$  n'est pas *identifié*, asymptotiquement ou autrement, par la procédure NLS. Par conséquent, une stratégie d'estimation assez différente, qui est en fait une stratégie essentiellement ad hoc, a dû être utilisée. Une conséquence malheureuse de cette méthode ad hoc est qu'aucune estimation de la variance de  $s^2$  n'est automatiquement disponible, rendant impossible une quelconque inférence statistique sur  $\sigma^2$ . Pour rendre une inférence possible, il faudrait connaître ou être capable d'estimer le quatrième moment des perturbations  $u_t$ , comme nous le montrons dans ce qui suit. Nous construisons, par analogie avec les résultats de l'estimateur  $\hat{\beta}$ , la variable aléatoire  $n^{1/2}(s^2 - \sigma_0^2)$ . A partir de (5.64) et des arguments qui la suivent, nous concluons que  $\hat{\mathbf{u}}^\top \hat{\mathbf{u}} = \mathbf{u}^\top \mathbf{u} + O(1)$ . Ainsi,

nous pouvons écrire

$$n^{1/2}(s^2 - \sigma_0^2) = n^{-1/2} \sum_{t=1}^n (u_t^2 - \sigma_0^2) + O(n^{-1/2}).$$

Puisqu'il est possible d'appliquer immédiatement un théorème de la limite centrale au premier terme du membre de droite de cette relation, nous pouvons de fait conclure que  $s^2$  est asymptotiquement normale. Mais la variance asymptotique est simplement  $\text{Var}(u_t^2)$ , et pour la calculer nous aurions besoin de connaître le quatrième moment de  $u_t$ .

Le quatrième moment de  $u_t$  pourrait évidemment être estimé, mais nous n'entamerons cette discussion ici. Le point important pour l'instant est que  $\sigma^2$  n'est pas un paramètre du modèle de régression non linéaire au sens habituel, et il faut par conséquent quitter le contexte de régression non linéaire si l'on souhaite réaliser une inférence statistique sur  $\sigma^2$ . Comme il existe d'autres méthodes d'estimation que les NLS, pour lesquelles  $\sigma^2$  n'a pas un statut particulier, nous devons conserver ce point à l'esprit lorsque nous travaillons avec les NLS. L'une de ces autres méthodes est bien évidemment celle du maximum de vraisemblance, et lorsque nous serons amenés dans les Chapitres 8 et 9 à examiner les modèles de régression à travers le prisme de cette méthode, nous verrons que des hypothèses sur la distribution des aléas, très différentes de celles posée dans le contexte purement NLS, seront nécessaires.

## 5.7 TESTS BASÉS SUR DES ESTIMATIONS NLS

Dans cette section, nous démontrons certains résultats de la Section 3.6 concernant les distributions asymptotiques des statistiques de test basées sur les estimations NLS. La plupart des résultats se révèlent être des conséquences directes de la normalité asymptotique de l'estimateur NLS, de la forme (5.25) de sa matrice de covariance limite, et de la convergence de  $\hat{\sigma}^2$ .

Nous généraliserons légèrement le traitement proposé dans la Section 3.6 en considérant un ensemble de restrictions *non linéaires* sur les paramètres  $\beta$  d'un modèle de régression non linéaire. Ces restrictions peuvent s'écrire comme

$$\mathbf{r}(\beta) = \mathbf{0}, \quad (5.65)$$

où le nombre de restrictions est  $r$  ( $< k$ ), et l'application  $\mathbf{r}$  (au moins) deux fois continûment différentiable part de l'espace paramétrique  $\Theta$  vers  $\mathbb{R}^r$ . Nous noterons  $\mathbf{R}(\beta)$  la matrice Jacobienne de dimension  $r \times k$  de  $\mathbf{r}(\beta)$  et supposerons que cette matrice est de plein rang  $r$ . Si elle ne l'est pas, soit certaines restrictions seraient redondantes, soit l'ensemble des restrictions serait impossible à satisfaire. Si  $\mathbf{R}(\beta)$  est évaluée en  $\beta_0$ , le vecteur paramétrique qui correspond au DGP, nous la noterons comme d'habitude  $\mathbf{R}_0$  à la place de  $\mathbf{R}(\beta_0)$ . Puisque nous supposons que le DGP satisfait les restrictions,  $\mathbf{r}_0 \equiv \mathbf{r}(\beta_0) = \mathbf{0}$ .

La statistique de test la plus facile à traiter est la statistique de Wald, l'expression (3.41). Pour les restrictions non linéaires (5.65), elle s'écrit

$$\frac{1}{\hat{\sigma}^2} \hat{\mathbf{r}}^\top (\hat{\mathbf{R}} (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{R}}^\top)^{-1} \hat{\mathbf{r}}, \quad (5.66)$$

où, comme d'habitude,  $\hat{\mathbf{r}} \equiv \mathbf{r}(\hat{\boldsymbol{\beta}})$ ,  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\boldsymbol{\beta}})$ , et  $\hat{\mathbf{R}} \equiv \mathbf{R}(\hat{\boldsymbol{\beta}})$ . Nous devons à présent opérer un développement de Taylor autour de  $\boldsymbol{\beta}_0$  des éléments qui apparaissent dans (5.66). Nous traiterons tout d'abord un seul élément de la fonction vectorielle  $\mathbf{r}(\boldsymbol{\beta})$ , comme suit :

$$r_i(\hat{\boldsymbol{\beta}}) = r_i(\boldsymbol{\beta}_0) + \mathbf{R}_i(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (D^2 r_i(\boldsymbol{\beta}^*)) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \quad (5.67)$$

où  $\mathbf{R}_i$  désigne la  $i^{\text{ième}}$  ligne de la matrice Jacobienne  $\mathbf{R}$ , et  $D^2 r_i$  désigne la matrice Hessienne de dimension  $k \times k$  de  $r_i$ . Comme d'habitude dans un développement de Taylor,  $\boldsymbol{\beta}^*$  est une combinaison convexe quelconque de  $\hat{\boldsymbol{\beta}}$  et  $\boldsymbol{\beta}_0$ . Comme  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O(n^{-1/2})$ , les deuxième et troisième termes du membre de droite de (5.67) sont  $O(n^{-1/2})$  et  $O(n^{-1})$ , respectivement. Puisque  $r_i(\boldsymbol{\beta}_0)$  est nul d'après (5.65), nous pouvons multiplier (5.67) par  $n^{1/2}$  pour obtenir une équation dans laquelle les termes d'ordre dominant sont  $O(1)$ . Si nous ne travaillons qu'avec l'ordre dominant, nous pouvons traiter le vecteur  $\mathbf{r}$  immédiatement et obtenir

$$n^{1/2} \mathbf{r}(\hat{\boldsymbol{\beta}}) = \mathbf{R}_0 n^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + O(n^{-1/2}). \quad (5.68)$$

Le premier terme du membre de droite de (5.68) est juste une linéarisation de  $\mathbf{r}(\hat{\boldsymbol{\beta}})$ . Si les restrictions étaient linéaires, (5.68) serait vraie sans le terme  $O(n^{-1/2})$ . Le type de résultat illustré dans (5.68) survient très fréquemment. Le fait que  $\mathbf{r}(\boldsymbol{\beta})$  soit *deux fois* continûment différentiable signifie que le Théorème de Taylor peut s'appliquer à l'ordre deux, et il est alors possible de déterminer à partir du dernier terme de ce développement l'ordre exact de l'erreur, dans ce cas  $O(n^{-1})$ , commise en négligeant ce terme. Par la suite, nous ne serons pas explicites sur ce raisonnement et mentionnerons simplement que le fait d'être deux fois continûment différentiable produit un résultat similaire à (5.68).

Les quantités dans (5.66) autres que  $\hat{\mathbf{r}}$  sont **asymptotiquement non stochastiques**. Nous entendons par là que

$$\hat{\mathbf{R}} = \mathbf{R}_0 + O(n^{-1/2}) \quad \text{et} \quad \hat{\mathbf{X}} = \mathbf{X}_0 + O(n^{-1/2}). \quad (5.69)$$

A nouveau, un développement en série de Taylor à l'ordre un cette fois conduit à ces résultats. Ils sont censés être interprétés élément par élément pour les matrices  $\mathbf{R}$  et  $\mathbf{X}$ . Cela ne porte pas à conséquence pour la matrice  $\mathbf{R}$  de dimension  $r \times k$ , mais cela est important pour la matrice  $\mathbf{X}$  de dimension  $n \times k$ . Nous devons être vigilants parce que dans les produits matriciels comme

$\hat{\mathbf{X}}^\top \hat{\mathbf{X}}$ , nous manipulons des sommes de  $n$  termes, qui seront évidemment d'ordres différents de ceux des termes des sommes en général. Cependant, si nous utilisons explicitement le fait que  $\hat{\mathbf{r}} = O(n^{-1/2})$  pour récrire (5.66) comme

$$(n^{1/2}\hat{\mathbf{r}})^\top (\hat{\sigma}^2 \hat{\mathbf{R}} (n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{R}}^\top)^{-1} (n^{1/2}\hat{\mathbf{r}}), \quad (5.70)$$

nous voyons qu'il faut gérer non pas  $\hat{\mathbf{X}}^\top \hat{\mathbf{X}}$  mais plutôt  $n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}}$ , et ce dernier *est* asymptotiquement non stochastique:

$$\begin{aligned} n^{-1}(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})_{ij} &= n^{-1} \sum_{t=1}^n \hat{X}_{ti} \hat{X}_{tj} \\ &= n^{-1} \sum_{t=1}^n (X_{ti}^0 + O(n^{-1/2})) (X_{tj}^0 + O(n^{-1/2})) \\ &= n^{-1} \sum_{t=1}^n (X_{ti}^0 X_{tj}^0 + O(n^{-1/2})) \\ &= n^{-1} (\mathbf{X}_0^\top \mathbf{X}_0)_{ij} + O(n^{-1/2}), \end{aligned}$$

où  $X_{ti}^0$  désigne le  $ti^{\text{ième}}$  élément de  $\mathbf{X}_0$ . La deuxième ligne utilise (5.69). La troisième ligne en découle parce que la somme de  $n$  termes d'ordre  $n^{-1/2}$  peut être au plus d'ordre  $n^{1/2}$ ; divisée par  $n$ , elle devient d'ordre  $n^{-1/2}$ . Notons que  $n^{-1} \mathbf{X}_0^\top \mathbf{X}_0$  lui-même est  $O(1)$ .

Puis, nous utilisons le résultat de la normalité asymptotique (5.39) pour obtenir une expression plus pratique pour  $n^{1/2}\hat{\mathbf{r}}$ . Nous avons

$$n^{-1/2}\hat{\mathbf{r}} = \mathbf{R}_0 (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u} + o(1). \quad (5.71)$$

Si nous substituons (5.71) dans (5.70), exploitons le déterminisme asymptotique de  $\hat{\mathbf{R}}$  et  $n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}}$ , et utilisons le fait que  $\hat{\sigma}^2$  tend en probabilité vers  $\sigma_0^2$ , la statistique de Wald est asymptotiquement équivalente à

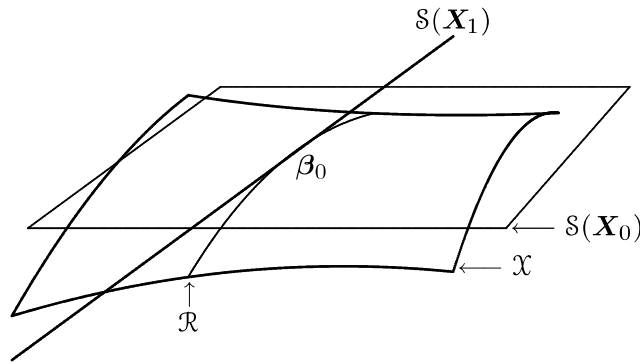
$$\sigma_0^{-2} \mathbf{u}^\top \mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{u}.$$

Il est facile de constater que cette expression est simplement

$$\sigma_0^{-2} \mathbf{u}^\top \mathbf{P}_2 \mathbf{u}, \quad (5.72)$$

où  $\mathbf{P}_2$  est la projection orthogonale sur l'espace engendré par les  $r$  colonnes de la matrice  $\mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{R}_0^\top$ . Cette projection orthogonale possède une interprétation géométrique et statistique très intéressante, que nous exposons à présent. Elle justifiera le choix apparemment étrange de l'indice 2 dans  $\mathbf{P}_2$ .

Considérons tout d'abord le sous-espace linéaire  $\mathcal{S}(\mathbf{X}_0)$ , qui est l'espace d'arrivée de la projection  $\mathbf{P}_0$ . Ce sous-espace est de dimension  $k$ , la dimension de l'espace paramétrique entier non contraint  $\Theta$ , puisqu'il est tangent



**Figure 5.1** Le sous-espace linéaire  $\mathcal{S}(\mathbf{X}_1)$  de  $\mathcal{S}(\mathbf{X}_0)$

à la variété incurvée de dimension  $k$   $\mathcal{X}$  générée par la variation du vecteur paramétrique  $\boldsymbol{\beta}$  de dimension  $k$  au point  $\mathcal{X}(\boldsymbol{\beta}_0)$ . (Consulter la discussion de la Figure 2.2 pour un rappel sur cette notation.)

Nous pouvons définir une sous-variété  $\mathcal{R}$  de  $\mathcal{X}$ , de dimension  $k - r$ , en contraignant la variation de  $\boldsymbol{\beta}$  à des valeurs qui satisfont les restrictions (5.65). En particulier, le point  $\mathcal{X}(\boldsymbol{\beta}_0)$  appartient à  $\mathcal{R}$  parce que nous avons supposé que  $\boldsymbol{\beta}_0$  satisfait les restrictions. Cette sous-variété, comme  $\mathcal{X}$  elle-même, a un espace tangent en  $\mathcal{X}(\boldsymbol{\beta}_0)$ , qui est un sous-espace (linéaire) de l'espace tangent entier  $\mathcal{S}(\mathbf{X}_0)$ . Nous noterons  $\mathcal{S}(\mathbf{X}_1)$  cet espace tangent contraint et  $\mathbf{P}_1$  la projection orthogonale qui lui est associée.<sup>6</sup> Les variétés  $\mathcal{X}$  et  $\mathcal{R}$ , ainsi que les espaces tangents  $\mathcal{S}(\mathbf{X}_0)$  et  $\mathcal{S}(\mathbf{X}_1)$ , sont illustrés dans la Figure 5.1.

Algébriquement, l'espace tangent  $\mathcal{S}(\mathbf{X}_0)$  peut être caractérisé comme l'ensemble de toutes les combinaisons linéaires des colonnes de la matrice  $\mathbf{X}_0$ . Tous les vecteurs du sous-espace  $\mathcal{S}(\mathbf{X}_1)$  sont nécessairement de telles combinaisons linéaires. Supposons ensuite que pour un vecteur  $\mathbf{b}$  quelconque de dimension  $k \times 1$  le vecteur  $\mathbf{X}_0 \mathbf{b}$  appartienne à  $\mathcal{S}(\mathbf{X}_1)$ . Nous montrons maintenant que ceci est le cas si et seulement si  $\mathbf{b}$  satisfait la relation  $\mathbf{R}_0 \mathbf{b} = \mathbf{0}$ .

Supposons que  $\boldsymbol{\beta}_1$  obéisse aux restrictions (5.65) et soit proche de  $\boldsymbol{\beta}_0$ . Alors, à partir d'un développement en série de Taylor,

$$\mathbf{x}(\boldsymbol{\beta}_1) = \mathbf{x}(\boldsymbol{\beta}_0) + \mathbf{X}^*(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0), \quad (5.73)$$

où  $\mathbf{X}^* \equiv \mathbf{X}(\boldsymbol{\beta}^*)$  et, comme d'habitude,  $\boldsymbol{\beta}^*$  est une combinaison convexe de  $\boldsymbol{\beta}_0$  et  $\boldsymbol{\beta}_1$ . Si  $\boldsymbol{\beta}_1$  converge vers  $\boldsymbol{\beta}_0$  par des valeurs qui satisfont toujours (5.65), alors la tangente en  $\mathcal{X}(\boldsymbol{\beta}_0)$  à la courbe le long de laquelle  $\boldsymbol{\beta}_1$  se dirige vers  $\boldsymbol{\beta}_0$  est la limite, quand  $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\| \rightarrow \mathbf{0}$ , du vecteur de dimension  $n$

$$\frac{\mathbf{x}(\boldsymbol{\beta}_1) - \mathbf{x}(\boldsymbol{\beta}_0)}{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\|}.$$

<sup>6</sup> Notre notation diffère ici de celle qui est souvent utilisée en connexion avec les tests d'hypothèses. Il n'est pas rare de noter  $H_0$  l'hypothèse nulle et  $H_1$  l'hypothèse. Si cette convention était adoptée, tous les indices 0 et 1 seraient intervertis.

D'après (5.73), cette limite est simplement  $\mathbf{X}_0 \mathbf{b}$ , où  $\mathbf{b}$  est défini comme le vecteur de dimension  $k$  qui est la limite de  $(\beta_1 - \beta_0)/\|\beta_1 - \beta_0\|$  quand  $\beta_1$  tend vers  $\beta_0$ . Ainsi,  $\mathbf{b}$  est simplement la limite d'un vecteur unité dans la direction du segment de droite qui relie  $\beta_0$  à  $\beta_1$ .

Puisque  $\mathbf{r}(\beta_1) = \mathbf{0}$ , un autre développement de Taylor révèle que

$$\mathbf{0} = \mathbf{R}(\beta^*)(\beta_1 - \beta_0).$$

Si  $\beta_1$  tend vers  $\beta_0$  comme précédemment, un calcul exactement similaire au précédent montre que  $\mathbf{R}_0 \mathbf{b} = \mathbf{0}$ . Ainsi, des tangentes à toutes les courbes qui appartiennent à  $\mathcal{R}$  et qui passent par  $\mathcal{X}(\beta_0)$  peuvent s'exprimer en terme de  $\mathbf{X}_0 \mathbf{b}$  pour un vecteur  $\mathbf{b}$  de dimension  $k$  qui satisfait  $\mathbf{R}_0 \mathbf{b} = \mathbf{0}$ . Nous pouvons facilement vérifier que l'argument qui vient d'être livré fonctionne aussi bien dans la direction opposée, et il s'ensuit que la condition  $\mathbf{R}_0 \mathbf{b} = \mathbf{0}$  est nécessaire et suffisante pour que  $\mathbf{X}_0 \mathbf{b}$  appartienne à  $\mathcal{S}(\mathbf{X}_1)$ . Notons que  $\mathcal{S}(\mathbf{X}_1)$  peut aussi s'exprimer en terme de la projection  $\mathbf{P}_1$ , comme  $\mathcal{S}(\mathbf{P}_1)$ .

Si  $\mathbf{R}_0 \mathbf{b} = \mathbf{0}$ , le vecteur  $\mathbf{X}_0 \mathbf{b}$  est orthogonal à toutes les colonnes de la matrice

$$\mathbf{X}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{R}_0^\top$$

et par là même à chaque vecteur de  $\mathcal{S}(\mathbf{P}_2)$ . Ce résultat est immédiat puisque

$$\mathbf{R}_0 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{X}_0 \mathbf{b} = \mathbf{R}_0 \mathbf{b} = \mathbf{0}.$$

Ainsi, les deux sous-espaces  $\mathcal{S}(\mathbf{P}_1)$  et  $\mathcal{S}(\mathbf{P}_2)$  sont mutuellement orthogonaux. Ce sont deux sous-espaces de  $\mathcal{S}(\mathbf{X}_0)$ , de dimensions  $k - r$  et  $r$ , respectivement. Puisque  $\mathcal{S}(\mathbf{X}_0)$  est lui-même de dimension  $k$ , il s'ensuit que la projection orthogonale qui lui est associée est la somme des deux autres:

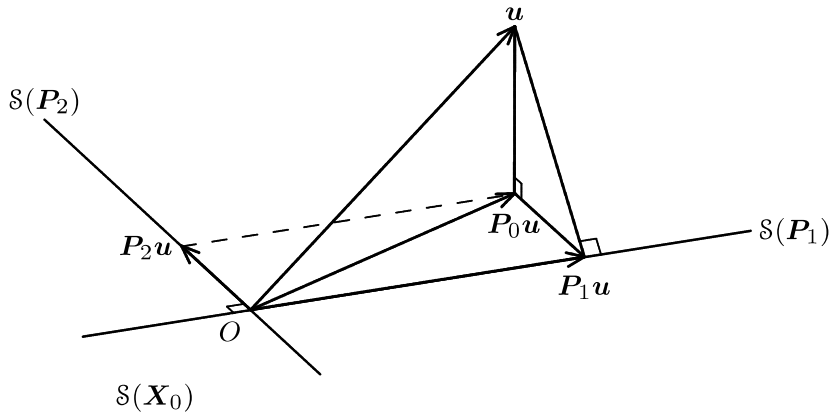
$$\mathbf{P}_0 = \mathbf{P}_1 + \mathbf{P}_2.$$

En utilisant ce résultat et (5.72), nous pouvons obtenir une autre expression pour la variable aléatoire vers laquelle tend asymptotiquement la statistique de Wald:

$$\begin{aligned} \sigma_0^{-2} \mathbf{u}^\top \mathbf{P}_2 \mathbf{u} &= \sigma_0^{-2} \mathbf{u}^\top (\mathbf{P}_0 - \mathbf{P}_1) \mathbf{u} \\ &= \sigma_0^{-2} \|(\mathbf{P}_0 - \mathbf{P}_1) \mathbf{u}\|^2 = \sigma_0^{-2} \|\mathbf{P}_0 \mathbf{u} - \mathbf{P}_1 \mathbf{u}\|^2. \end{aligned} \quad (5.74)$$

Ce résultat, illustré dans la Figure 5.2, est riche d'enseignements. Le vecteur  $\mathbf{P}_0 \mathbf{u}$  est la projection du vecteur d'aléas  $\mathbf{u}$  sur  $\mathcal{S}(\mathbf{X}_0)$ . C'est  $\mathbf{P}_0 \mathbf{u}$  qui provoque l'erreur d'estimation dans les estimations NLS  $\hat{\beta}$ : si  $\mathbf{P}_0 \mathbf{u} = \mathbf{0}$ , nous aurions  $\hat{\beta} = \beta_0$  et par conséquent aucune erreur d'estimation. Le vecteur  $\mathbf{P}_1 \mathbf{u}$  est la projection de  $\mathbf{u}$ , mais également celle de  $\mathbf{P}_0 \mathbf{u}$ , sur  $\mathcal{S}(\mathbf{P}_1)$ , le sous-espace de dimension  $k - r$  de  $\mathcal{S}(\mathbf{X}_0)$  qui correspond aux restrictions. Ce vecteur est la source de la partie de l'erreur d'estimation qui n'enfreint pas les restrictions (5.65). L'autre partie de  $\mathbf{P}_0 \mathbf{u}$  est la différence  $(\mathbf{P}_0 - \mathbf{P}_1) \mathbf{u} = \mathbf{P}_2 \mathbf{u}$ , orthogonale





**Figure 5.2** Les projections qui conduisent au test de Wald

au sous-espace  $\mathcal{S}(\mathbf{X}_0)$ . Cette partie de l'erreur d'estimation conduit à une estimation non contrainte qui en général ne satisfait pas (5.65). La variable aléatoire (5.74) s'interprète comme la longueur au carré de cette composante de l'erreur d'estimation, normalisée par la variance des  $u_t$ . À condition que la véritable valeur  $\beta_0$  satisfasse les restrictions (5.65), la variable (5.74) devrait par conséquent posséder une distribution du chi-deux à  $r$  degrés de liberté. Cependant, si  $\mathbf{x}(\beta_0)$  n'appartient pas à la variété contrainte  $\mathcal{R}$ , la variable comprendra un terme non aléatoire correspondant au carré de la distance entre  $\mathbf{x}(\beta_0)$  et  $\mathcal{R}$  et sera par conséquent plus importante.

La deuxième statistique de test est le multiplicateur de Lagrange, ou statistique LM, qui est l'expression (3.47) dans sa forme LM et (3.48) dans sa forme du vecteur score. Puisque ces deux formes de la statistique LM sont numériquement identiques, nous nous focaliserons uniquement sur la dernière.

Notons tout d'abord que, par analogie avec (5.62), les résidus  $\tilde{\mathbf{u}} \equiv \mathbf{y} - \mathbf{x}(\tilde{\beta})$  à partir de l'estimation contrainte satisfont

$$\tilde{\mathbf{u}} \stackrel{a}{=} \mathbf{M}_1 \mathbf{u}, \quad (5.75)$$

puisque  $\mathbf{P}_1$  joue le même rôle pour la variété  $\mathcal{R}$  que  $\mathbf{P}_0$  pour  $\mathcal{X}$ . La statistique LM (3.48) est

$$\frac{1}{\tilde{\sigma}^2} (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{P}}_X (\mathbf{y} - \tilde{\mathbf{x}}). \quad (5.76)$$

Si nous l'exprimons en terme de quantités  $O(1)$ , nous obtenons

$$\frac{1}{\tilde{\sigma}^2} n^{-1/2} (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}} (n^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} n^{-1/2} \tilde{\mathbf{X}}^\top (\mathbf{y} - \tilde{\mathbf{x}}). \quad (5.77)$$

Comme  $\hat{\mathbf{X}}_t$ ,  $\tilde{\mathbf{X}}_t$  est asymptotiquement non stochastique. Ainsi, de (5.75),

$$\begin{aligned} n^{-1/2} \tilde{\mathbf{X}}^\top (\mathbf{y} - \tilde{\mathbf{x}}) &= n^{-1/2} \sum_{t=1}^n \tilde{\mathbf{X}}_t^\top \tilde{u}_t \\ &= n^{-1/2} \sum_{t=1}^n \mathbf{X}_{0t}^\top (\mathbf{M}_1 \mathbf{u})_t + o(1) \\ &= n^{-1/2} \sum_{t=1}^n (\mathbf{M}_1 \mathbf{X}_0)_t u_t + o(1) \\ &= n^{-1/2} \mathbf{X}_0^\top \mathbf{M}_1 \mathbf{u} + o(1). \end{aligned}$$

La matrice  $n^{-1} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$  est asymptotiquement non stochastique, tout comme l'est  $n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}}$ , de sorte que la statistique LM (5.77) est asymptotiquement équivalente à

$$\mathbf{u}^\top \mathbf{M}_1 \mathbf{X}_0 (\sigma_0^2 \mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0 \mathbf{M}_1 \mathbf{u} = \sigma_0^{-2} \mathbf{u}^\top \mathbf{M}_1 \mathbf{P}_0 \mathbf{M}_1 \mathbf{u}. \quad (5.78)$$

Puisque  $\mathcal{S}(\mathbf{X}_1)$  est un sous-espace de  $\mathcal{S}(\mathbf{X}_0)$ , nous avons  $\mathbf{P}_1 \mathbf{P}_0 = \mathbf{P}_0 \mathbf{P}_1 = \mathbf{P}_1$ , d'où il découle que  $\mathbf{M}_1 \mathbf{P}_0 \mathbf{M}_1 = \mathbf{P}_0 - \mathbf{P}_1$ . Ainsi, l'expression (5.78) devient

$$\sigma_0^{-2} \mathbf{u}^\top (\mathbf{P}_0 - \mathbf{P}_1) \mathbf{u} = \sigma_0^{-2} \mathbf{u}^\top \mathbf{P}_2 \mathbf{u}. \quad (5.79)$$

La comparaison entre (5.79) et (5.72) montre que la statistique LM est asymptotiquement égale à la statistique de Wald. Ainsi, elle est elle aussi asymptotiquement distribuée suivant une  $\chi^2(r)$  sous l'hypothèse nulle.

La dernière des trois statistiques de test discutée dans la Section 3.6 était basée sur le principe du rapport de vraisemblance, la statistique pseudo- $F$  (3.50). Puisque seuls les résultats asymptotiques nous intéressent, nous la récrivons ici sous une forme qui lui permet d'être asymptotiquement distribuée suivant une  $\chi^2(r)$ :

$$\frac{1}{s^2} (SSR(\tilde{\beta}) - SSR(\hat{\beta})) \quad (5.80)$$

et nous nous y référerons (de façon quelque peu imprécise) en tant que statistique LR. Nous avons déjà vu que  $s^2 \rightarrow \sigma_0^2$  quand  $n \rightarrow \infty$ . Il reste à montrer que  $SSR(\tilde{\beta}) - SSR(\hat{\beta})$ , divisée par  $\sigma_0^2$ , suit asymptotiquement une  $\chi^2(r)$ . A partir de (5.64), nous avons

$$\hat{\sigma}^2 = \frac{1}{n} \mathbf{u}^\top \mathbf{M}_0 \mathbf{u} + o(n^{-1}),$$

d'où nous obtenons, après une multiplication par  $n$ ,

$$SSR(\hat{\beta}) = \mathbf{u}^\top \mathbf{M}_0 \mathbf{u} + o(1).$$

Le résultat analogue pour la somme des carrés contrainte est

$$SSR(\tilde{\beta}) = \mathbf{u}^\top \mathbf{M}_1 \mathbf{u} + o(1),$$

de sorte que, à l'ordre dominant asymptotiquement, l'expression (5.80) devient

$$\sigma_0^{-2} \mathbf{u}^\top (\mathbf{M}_1 - \mathbf{M}_0) \mathbf{u} = \sigma_0^{-2} \mathbf{u}^\top (\mathbf{P}_0 - \mathbf{P}_1) \mathbf{u} = \sigma_0^{-2} \mathbf{u}^\top \mathbf{P}_2 \mathbf{u}. \quad (5.81)$$

L'expression la plus à droite dans (5.81) est précisément la variable aléatoire du membre de droite de (5.79) et la première expression dans (5.74), dont nous avons montré qu'elles étaient asymptotiquement équivalentes à la statistique LM et à la statistique de Wald, respectivement. Ainsi, nous concluons non seulement que la statistique (5.80) est asymptotiquement distribuée suivant une  $\chi^2(r)$  mais aussi qu'il s'agit asymptotiquement de la même variable aléatoire que les deux autres statistiques de test.

Nous pouvons donc réunir les résultats de cette section dans un théorème:

*Théorème 5.4.*

Pour un modèle de régression non linéaire (5.08) soumis aux restrictions (5.65), où à la fois les estimations contraintes  $\tilde{\beta}$  et les estimations non contraintes  $\hat{\beta}$  convergent et sont asymptotiquement normales, la statistique de Wald (5.66), la statistique LM (5.76), et la statistique de test LR (5.80) sont, sous l'hypothèse nulle, asymptotiquement égales à la variable aléatoire

$$\sigma_0^{-2} \mathbf{u}^\top \mathbf{P}_2 \mathbf{u},$$

qui suit asymptotiquement une  $\chi^2(r)$ . Ici,  $\mathbf{P}_2 \equiv \mathbf{P}_0 - \mathbf{P}_1$ , où  $\mathbf{P}_0$  désigne la projection sur le sous-espace de dimension  $k$   $\mathcal{S}(\mathbf{X}_0)$ , et  $\mathbf{P}_1$  la projection sur le sous-espace de dimension  $k - r$  de  $\mathcal{S}(\mathbf{X}_0)$  qui correspond aux variations paramétriques satisfaisant les restrictions.

## 5.8 LECTURES COMPLÉMENTAIRES ET CONCLUSION

Nous avons fourni dans ce chapitre un traitement asymptotique relativement complet de l'estimation des modèles de régression non linéaire au moyen des moindres carrés non linéaires. Les lecteurs à la recherche d'un traitement encore plus complet, plus rigoureux, ou basé sur des hypothèses moins contraignantes sont orientés vers Jennrich (1969), qui est un article classique sur les propriétés asymptotiques de l'estimateur NLS, Malinvaud (1970b), Wu (1981), ou vers les ouvrages de White (1984), Gallant (1987), et Gallant et White (1988). Ces derniers développent la théorie asymptotique pour une grande variété de modèles qui intéressent les économètres. Des références légèrement moins techniques sont Amemiya (1983), Bates et Watts (1988), et Seber et Wild (1989, Chapitre 12). L'analyse de ce chapitre dépend dans

une large mesure du fait que l'estimateur NLS est défini par la minimisation de la fonction somme-des-carrés. Il s'avère que l'analyse s'applique pour de nombreux aspects à d'autres estimateurs définis par la minimisation ou la maximisation d'autres fonctions critères; consulter le Chapitre 17. Des traitements qui gèrent de façon abstraite les estimateurs définis de la sorte sont disponibles chez Amemiya (1985, Chapitre 4) et Huber (1981). Quand nous traiterons l'estimation par maximum de vraisemblance dans le Chapitre 8, les résultats de ce chapitre serviront de modèles à la dérivation de résultats similaires dans un autre contexte.

## TERMES ET CONCEPTS

application définissante des paramètres	estimation (des paramètres d'un modèle)
asymptotiquement non stochastique	identification asymptotique
caractérisation des DGP, complète ou partielle	identification asymptotique stricte
convergence au taux $n^{1/2}$	identification (d'un modèle paramétrisé)
convergence des estimateurs	innovations
développement d'un DGP à des échantillons arbitrairement grands	matrice de covariance asymptotique
développement stochastique	meilleur estimateur linéaire sans biaisé (BLUE)
dimension	modèle paramétrisé
distribution asymptotique	modèles dynamiques
efficacité asymptotique	normalité asymptotique
efficacité (d'un estimateur)	paramètres du modèle
espace d'arrivée (d'une application)	précision des estimateurs
espace de départ (d'une application)	processus explosif
espace paramétrique	règles pour la génération de processus stochastiques infinis
estimateur convergent et asymptotiquement linéaire	résidus NLS
estimateur (des paramètres d'un modèle)	Théorème de Gauss-Markov
estimateur et estimation	variables dépendantes retardées
	variables strictement exogènes

# Chapitre 6

## La Régression de Gauss-Newton

### 6.1 INTRODUCTION

On associe au modèle de régression non linéaire une **régression artificielle** que l'on nomme **régression de Gauss-Newton**, ou **GNR**. Nous avons déjà rencontré une version de la régression de Gauss-Newton; nous l'avons employée dans la Section 3.6 pour calculer les tests du multiplicateur de Lagrange pour les modèles de régression non linéaire. Les régressions artificielles sont tout simplement des régressions linéaires qui sont employées comme système d'évaluation. Comme nous le constaterons, de nombreux types de modèles non linéaires en économétrie possèdent des régressions artificielles associées. La régressande et les régresseurs sont construits délibérément de manière à ce que lorsque l'on exécute la régression artificielle, certaines valeurs qu'affiche le programme de régression sont des quantités que l'on désire calculer. De nombreux résultats donnés par les régressions artificielles peuvent pourtant n'être d'aucun intérêt. Par exemple, il nous arrive d'exécuter des régressions artificielles pour lesquelles toutes les estimations des coefficients seront nulles!

On emploie les régressions artificielles pour au moins cinq raisons différentes:

- (i) pour vérifier que les conditions du premier ordre d'un minimum ou d'un maximum sont satisfaites avec assez de précision;
- (ii) pour calculer les matrices de covariance estimées;
- (iii) pour calculer les statistiques de test après qu'un modèle ait été estimé sous contraintes, sans avoir besoin d'estimer le modèle non contraint;
- (iv) pour calculer des estimations efficaces en une étape;
- (v) c'est un élément essentiel pour les procédures d'optimisation numériques dont on fait usage pour chercher les estimations par moindres carrés non linéaires et d'autres types d'estimations.

Dans le présent chapitre nous discuterons de la manière de faire usage de la régression de Gauss-Newton pour ces différentes raisons. Puis, lorsque nous rencontrerons d'autres régressions artificielles, nous verrons qu'elles peuvent souvent être utilisées exactement de la même manière que la GNR. De fait, de nombreux résultats que nous obtiendrons dans ce chapitre réapparaîtront

à plusieurs reprises dans les chapitres qui suivront, mais des vecteurs et des matrices différentes remplaceront ceux que nous utiliserons ici. L'écriture algébrique (et son interprétation géométrique) sera identique dans tous les cas; seul le modèle statistique sous-jacent, et par là les définitions de la régressande et des régresseurs, seront modifiés.

Ainsi que nous l'avons fait dans les chapitres qui précèdent, nous traiterons le modèle de régression non linéaire univariée

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (6.01)$$

où  $\mathbf{X}(\boldsymbol{\beta}) \equiv D\mathbf{x}(\boldsymbol{\beta})$  est une matrice de dimension  $n \times k$  dont le  $t^{\text{ième}}$  élément est la dérivée partielle de  $x_t(\boldsymbol{\beta})$  par rapport à  $\beta_i$ . Le vecteur des fonctions,  $\mathbf{x}(\boldsymbol{\beta})$ , et sa matrice de dérivées,  $\mathbf{X}(\boldsymbol{\beta})$ , sont supposés satisfaire les conditions de convergence et de normalité asymptotique que nous avons détaillées dans les Théorèmes 5.1 et 5.2.

La façon la plus évidente d'obtenir la régression de Gauss-Newton consiste à calculer le développement en série de Taylor à l'ordre un de (6.01) autour d'un quelconque vecteur de paramètres  $\boldsymbol{\beta}^*$ . Cela donne

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}^*) + \mathbf{X}(\boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \text{termes de plus haut degré} + \mathbf{u}.$$

En remplaçant  $\mathbf{x}(\boldsymbol{\beta}^*)$  dans le membre de gauche, en combinant les termes de plus haut degré avec les aléas  $\mathbf{u}$  et en nommant tout ceci "résidus", et en remplaçant  $\boldsymbol{\beta} - \boldsymbol{\beta}^*$  par un vecteur  $(\mathbf{b})$  à  $k$  composantes que nous ne spécifions pas, nous obtenons

$$\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^*) = \mathbf{X}(\boldsymbol{\beta}^*)\mathbf{b} + \text{résidus}. \quad (6.02)$$

C'est la version la plus simple de la régression de Gauss-Newton dans sa forme générique. La régressande est comparable à un vecteur de résidus, car c'est la différence entre le vecteur des valeurs de la véritable variable dépendante et le vecteur de valeurs "prévues" par le modèle  $\mathbf{x}(\boldsymbol{\beta}^*)$ . Il y a  $k$  régresseurs, chacun étant un vecteur de dérivées de  $\mathbf{x}(\boldsymbol{\beta})$  par rapport à un élément de  $\boldsymbol{\beta}$ . Par conséquent cela a un sens d'associer le  $i^{\text{ième}}$  régresseur à  $\beta_i$ . Ainsi que nous l'avons vu, lorsque  $\mathbf{x}(\boldsymbol{\beta})$  est un modèle de régression linéaire où  $\mathbf{X}$  est la matrice des variables indépendantes,  $\mathbf{X}(\boldsymbol{\beta})$  est tout simplement égal à  $\mathbf{X}$ , de sorte que pour les modèles linéaires, la GNR possédera justement les mêmes régresseurs que le modèle d'origine.

Les propriétés de la régression (6.02) dépendront de la manière dont le vecteur de paramètres  $\boldsymbol{\beta}^*$  est choisi. Observons tout d'abord ce qu'il advient lorsque  $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}$ , l'estimation non contrainte NLS de  $\boldsymbol{\beta}$ . La régression de Gauss-Newton devient

$$\mathbf{y} - \hat{\mathbf{x}} = \hat{\mathbf{X}}\mathbf{b} + \text{résidus}, \quad (6.03)$$

où  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\boldsymbol{\beta}})$  et  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\boldsymbol{\beta}})$ . Maintenant rappelons-nous que les conditions du premier ordre pour un minimum de la fonction somme-des-carrés sont

$$(\mathbf{y} - \hat{\mathbf{x}})^\top \hat{\mathbf{X}} = \mathbf{0}. \quad (6.04)$$

L'estimation OLS de  $\mathbf{b}$  à partir de (6.03) est

$$\hat{\mathbf{b}} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top (\mathbf{y} - \hat{\mathbf{x}}),$$

qui doit être un vecteur nul si l'on tient compte de (6.04). Ainsi dans ce cas, la GNR n'aura strictement aucun pouvoir explicatif! Cela pourrait passer pour un résultat dénué d'intérêt. Après tout, pourquoi voudrait-on exécuter une régression artificielle dont on sait déjà par avance que tous les coefficients sont nuls? Il y a en réalité deux bonnes raisons de le faire.

La première est que la GNR (6.03) offre un moyen très simple de vérifier que les conditions du premier ordre (6.04) sont satisfaites avec suffisamment de précision. Etant données les limites de l'arithmétique à virgule flottante sur les calculateurs numériques, la *valeur approchée* de  $\hat{\boldsymbol{\beta}}$  qu'affiche un programme de moindres carrés non linéaires ne satisfera jamais réellement les conditions du premier ordre. Si le programme est performant et si les données procurent suffisamment d'informations pour nous permettre d'estimer  $\boldsymbol{\beta}$  avec précision, alors l'approximation  $\hat{\boldsymbol{\beta}}$  sera très proche du vrai  $\boldsymbol{\beta}$  et (6.04) sera presque vraie. En conséquence, l'estimation  $\hat{\mathbf{b}}$  de la GNR devrait être proche de zéro, et le pouvoir explicatif de la GNR sera presque nul. Dans de telles circonstances, on s'attend à ce que les Students de  $\hat{\mathbf{b}}$  affichés soient tous inférieurs à environ  $10^{-3}$  ou  $10^{-4}$ , et à ce que le  $R^2$  soit composé de plusieurs décimales nulles. Il est préférable d'observer les Students plutôt que  $\hat{\mathbf{b}}$  car les premiers sont des quantités sans unité de mesure; quelques éléments de  $\hat{\mathbf{b}}$  pourraient être assez importants, si les colonnes correspondantes de  $\hat{\mathbf{X}}$  étaient très petites, même si l'estimation de  $\hat{\boldsymbol{\beta}}$  était très fine.

S'il fallait exécuter la GNR (6.03) et si l'on s'apercevait que quelques Students associés à  $\hat{\mathbf{b}}$  étaient supérieurs à, disons  $10^{-2}$ , alors il faudrait mettre en doute la validité du  $\hat{\boldsymbol{\beta}}$  qui a été calculé. Peut-être le calcul devrait-il être exécuté à nouveau à l'aide d'un critère de convergence plus sévère ou d'un algorithme différent (consulter la Section 6.8). Ou alors les données et le modèle sont tels qu'une estimation plus fine de  $\hat{\boldsymbol{\beta}}$  est impossible ou très difficile à obtenir, auquel cas il serait souhaitable d'estimer un autre modèle, plus simple, ou de disposer de davantage de données.

La GNR (6.03) est particulièrement utile lorsqu'une évaluation de  $\hat{\boldsymbol{\beta}}$  a été calculée avec un programme de moindres carrés non linéaires auquel on ne peut pas se fier (de tels programmes existent!), ou grâce à une procédure ad hoc. Les procédures ad hoc sont parfois utilisées lorsque le modèle est seulement légèrement non linéaire. En particulier, il y a de très nombreux modèles non linéaires qui sont linéaires conditionnellement à un de leurs paramètres. Un exemple est le modèle

$$y_t = \beta_1 + \beta_2 z_{t1} + \beta_3 z_{t2}^{\beta_4} + u_t, \quad (6.05)$$

où les variables  $z_{t1}$  et  $z_{t2}$  sont des régresseurs exogènes, qui est linéaire conditionnellement à  $\beta_4$ . Il est parfois pratique d'estimer de tels modèles en faisant abstraction de l'unique paramètre qui engendre la non linéarité, dans ce cas,  $\beta_4$ , et de calculer les autres paramètres avec une procédure de moindres carrés ordinaires conditionnellement à chaque valeur de  $\beta_4$ . On exécute ensuite la GNR pour savoir si, oui ou non, l'approximation de  $\hat{\beta}$  qui en résulte est assez précise, ce qui, dans le cas de (6.05) donne

$$\begin{aligned} y_t - \hat{\beta}_1 - \hat{\beta}_2 z_{t1} - \hat{\beta}_3 z_{t2}^{\hat{\beta}_4} \\ = b_1 + b_2 z_{t1} + b_3 z_{t2}^{\hat{\beta}_4} + b_4 \hat{\beta}_3 (\log z_{t2}) z_{t2}^{\hat{\beta}_4} + \text{résidu.} \end{aligned} \quad (6.06)$$

Si les Students de  $\hat{b}_1$ ,  $\hat{b}_2$ ,  $\hat{b}_3$  et  $\hat{b}_4$  sont tous suffisamment faibles, on peut accepter le  $\hat{\beta}$  calculé comme étant suffisamment proche des estimations NLS.

Dans la Section 1.6 nous avons discuté des techniques de détection des points de levier et des observations influentes dans le contexte des OLS. Dans le cas d'un modèle de régression non linéaire, il est possible d'appliquer ces techniques à la GNR (6.03). Le  $t^{\text{ième}}$  élément diagonal de la matrice chapeau pour la GNR est

$$\hat{h}_t \equiv \hat{\mathbf{X}}_t (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}_t^\top,$$

que de nombreux progiciels de régression savent calculer assez facilement comme un sous-produit du calcul des estimations OLS. Avec  $\hat{h}_t$ , il est possible de calculer

$$\left( \frac{\hat{h}_t}{1 - \hat{h}_t} \right) \hat{u}_t$$

pour tout  $t$ . Cette expression est l'analogie de l'expression (1.42), et en dessiner la courbe représentative en fonction de  $t$  est un bon moyen de détecter les observations influentes. Evidemment, dans le cas non linéaire, cela ne sera pas exactement égal à la modification enregistrée par le résidu de l'observation  $t$  apportée par l'omission de cette observation dans la régression, contrairement à ce qui se passerait si  $\mathbf{x}(\beta)$  était linéaire en  $\beta$ , mais cela offrira néanmoins une bonne approximation dans la plupart des situations. Ainsi, en appliquant les techniques habituelles de détection des observations influentes dans les modèles de régression linéaire à la GNR (6.03), il est possible de détecter les problèmes relatifs aux données dans les modèles de régression non linéaire aussi simplement que pour les modèles linéaires.

## 6.2 CALCUL DES MATRICES DE COVARIANCE

La seconde raison majeure pour laquelle on utilise la GNR (6.03) est qu'elle permet le calcul de la matrice de covariance *estimée* de  $\hat{\beta}$ . Souvenons-nous



du résultat asymptotique, le Théorème 5.2, selon lequel pour un modèle de régression non linéaire correctement spécifié,

$$n^{1/2}(\hat{\beta} - \beta_0) \overset{a}{\sim} N(\mathbf{0}, \sigma_0^2(n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}), \quad (6.07)$$

où  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$ . Dans la pratique nous sommes intéressés par la distribution de  $\hat{\beta} - \beta_0$  plutôt que par celle de  $n^{1/2}(\hat{\beta} - \beta_0)$ , et pour obtenir une matrice de covariance estimée il nous faut d'abord remplacer  $\sigma_0^2$  et  $(n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}$  dans (6.07) par des quantités qui les estiment de manière convergente, et les diviser par  $n$ .

Considérons à présent la GNR (6.03) une nouvelle fois. L'estimation de la matrice de covariance que le programme de régression affichera est

$$s^2(\hat{\mathbf{X}}^\top\hat{\mathbf{X}})^{-1}, \quad (6.08)$$

où

$$s^2 \equiv \frac{(\mathbf{y} - \hat{\mathbf{x}})^\top(\mathbf{y} - \hat{\mathbf{x}})}{n - k}$$

est l'estimation OLS de la variance de la régression, à la fois de la régression artificielle et de la régression non linéaire originelle (6.01). Parce que la GNR n'a aucun pouvoir explicatif, ces deux régressions possèdent exactement les mêmes résidus.

Il est évident que  $s^2$  est une estimation convergente de  $\sigma_0^2$ , et puisque  $\hat{\beta}$  est une estimation convergente de  $\beta_0$ ,  $n^{-1}\hat{\mathbf{X}}^\top\hat{\mathbf{X}}$  doit être une estimation convergente de  $n^{-1}\mathbf{X}_0^\top\mathbf{X}_0$ ; voir Section 5.7. Ainsi, à l'évidence, il est raisonnable de faire usage de (6.08) pour estimer la matrice de covariance de  $\hat{\beta} - \beta_0$ . La matrice de covariance de  $\hat{\mathbf{b}}$  qu'affiche d'habitude le programme de moindres carrés offrira une estimation tout à fait correcte, et facilement calculable, de la matrice de covariance des estimations NLS. Particulièrement lorsque  $\hat{\beta}$  a été estimé par une tout autre méthode que celle des moindres carrés non linéaires (souvenons-nous du modèle (6.05) et de la GNR associée à (6.05)), la GNR (6.03) présente un moyen extrêmement simple de calculer une estimation de la matrice de covariance de  $\hat{\beta}$ .

Dans le cas non linéaire, il n'est pas évident qu'il faille utiliser  $s^2$  plutôt qu'un autre estimateur convergent de  $\sigma^2$ . L'estimateur le plus simple est

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \hat{\mathbf{x}})^\top(\mathbf{y} - \hat{\mathbf{x}}),$$

et, comme nous le verrons au cours du Chapitre 8, c'est l'estimateur que la méthode du maximum de vraisemblance conseille d'utiliser. Toutefois, comme nous l'avons montré à la Section 3.2,  $\hat{\sigma}^2$  tendra à sous-estimer  $\sigma^2$  en moyenne, de sorte qu'utiliser l'estimation OLS de (6.03) se révèle préférable. De plus, lorsque le DGP est un cas particulier du modèle que l'on estime, nous disposons du résultat asymptotique (5.57) selon lequel

$$\mathbf{y} - \hat{\mathbf{x}} \overset{a}{=} \mathbf{M}_0\mathbf{u}, \quad (6.09)$$

où  $M_0 \equiv \mathbf{I} - \mathbf{X}_0(\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top$  est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}_0)$ . Le résultat (6.09) est l'analogie du résultat valable en échantillon fini, pour les modèles de régression linéaire, qui est

$$\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = M_X \mathbf{u}.$$

Alors le fait que  $s^2$  soit un estimateur non biaisé de  $\sigma^2$  pour les modèles de régression linéaire suggère que, pour les échantillons de grande taille,  $s^2$  soit également non biaisé pour les modèles non linéaires et sûrement plus approprié que  $\hat{\sigma}^2$ .

Malheureusement, tous les programmes de moindres carrés non linéaires n'utilisent pas  $s^2$ . Une raison d'exécuter la GNR (6.03) est de voir si la matrice de covariance estimée pour  $\hat{\boldsymbol{\beta}}$  calculée par le programme est réellement la même que celle obtenue par la GNR. Les deux approches diffèrent du facteur de proportionnalité  $(n - k)/n$ , auquel cas il serait judicieux d'utiliser la plus grande des deux estimations (c'est-à-dire celle donnée par la GNR). Si elles diffèrent d'une autre façon, il serait souhaitable de se méfier du programme de moindres carrés non linéaires, et de conserver la matrice de covariance estimée par la GNR plutôt que celle affichée par le programme. Il y a cependant une exception possible à ce conseil. Certains progiciels modernes sont capables d'afficher une **estimation de la matrice de covariance robuste à l'hétéroscédasticité**, ou **HCCME**, au lieu de l'estimation usuelle (qui suppose l'homoscédasticité) dont nous venons de parler; voir Section 16.3. Dans des cas semblables, il serait sans doute révélateur d'exécuter la GNR et de comparer les deux estimations des matrices de covariance.

### 6.3 COLINÉARITÉ DANS LES RÉGRESSIONS NON LINÉAIRES

Nous avons remarqué dans la Section 2.3 que les modèles de régression linéaire qui sont insuffisamment identifiés sont souvent considérés comme manifestant un phénomène de **colinéarité** ou de **multicolinéarité**. Les deux termes expriment la même chose, aussi préférons-nous le plus court, même si le terme “multicolinéarité” est probablement le plus usité dans la littérature consacrée à l'économétrie. La relation entre la colinéarité et l'identification apparaît clairement si l'on étudie la régression de Gauss-Newton. Comme nous allons le montrer, si un modèle de régression non linéaire est faiblement identifié au voisinage d'un quelconque vecteur de paramètres  $\boldsymbol{\beta}^*$  qui n'est pas trop éloigné de  $\boldsymbol{\beta}_0$  (souvenons-nous que pour les modèles non linéaires, l'identification dépendra généralement des valeurs des paramètres), alors la GNR évaluée en  $\boldsymbol{\beta}^*$  manifestera la colinéarité. Nous présenterons également une explication intuitive de ce qu'implique la colinéarité dans les modèles de régression linéaire et non linéaire.

Dans la Section 2.3, nous avons défini sans trop de précision un **modèle insuffisamment identifié** comme un modèle pour lequel la matrice Hessienne

$\mathbf{H}(\boldsymbol{\beta})$  de la fonction somme-des-carrés est presque singulière pour des valeurs intéressantes de  $\boldsymbol{\beta}$ . Nous avons délibérément évité toute tentative d'approfondissement, car le fait qu'un modèle soit ou non insuffisamment identifié dépend de la raison pour laquelle nous nous soucions du fait qu'il est identifiable. Par exemple un modèle peut être tellement mal identifié que certains programmes de moindres carrés non linéaires soient incapables de l'estimer, mais pas suffisamment pour empêcher les meilleurs programmes de le manipuler et de le traiter. Ou bien un modèle peut être suffisamment bien identifié pour qu'il n'y ait pas de difficulté à l'estimer, mais pas assez pour permettre d'obtenir des estimations des paramètres aussi précises que nous le désirons.

Normalement, la quasi singularité d'une matrice est déterminée par l'observation de son **nombre condition** qui est le rapport de sa plus grande valeur propre par sa plus petite valeur propre. Cependant, le nombre condition de  $\mathbf{H}(\boldsymbol{\beta})$  peut être modifié grandement par une reparamétrisation du modèle, même si la reparamétrisation consiste simplement à choisir une autre échelle pour certains régresseurs. Par exemple, dans un cas de régression linéaire pour lequel  $\mathbf{x}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ , la matrice Hessienne  $\mathbf{H}(\boldsymbol{\beta})$  est égale à  $2\mathbf{X}^\top\mathbf{X}$ . Si l'on multiplie une colonne de  $\mathbf{X}$  par, disons,  $10^6$ , cela modifierait sérieusement le nombre condition de  $\mathbf{X}^\top\mathbf{X}$ , le rendant presque à coup sûr plus important, à moins que les éléments de cette colonne de  $\mathbf{X}$  ne fussent très petits à l'origine. Puisque ce genre de modification ne nous apprendra pas grand chose de plus à partir des données (bien qu'il puisse tout à fait affecter les performances des algorithmes de moindres carrés non linéaires), il ne nous est pas possible de classer raisonnablement les modèles selon leur degré d'identification sur la base d'un indice aussi simple que le nombre condition.<sup>1</sup> En réalité, il semble qu'il n'existe aucun procédé mécanique qui permette de décider si, oui ou non, un modèle est "mal" identifié. Toutefois, comme nous allons le voir à présent, la régression de Gauss-Newton peut être utilisée pour nous indiquer si l'identification peut être un obstacle.

Dans la Section 5.4 nous avons vu que

$$n^{-1}\mathbf{H}(\boldsymbol{\beta}_0) \stackrel{a}{=} 2n^{-1}\mathbf{X}^\top(\boldsymbol{\beta}_0)\mathbf{X}(\boldsymbol{\beta}_0). \quad (6.10)$$

Ainsi, si la matrice Hessienne est quasi singulière, nous pouvons nous attendre à ce qu'il en soit de même pour la matrice  $\mathbf{X}^\top(\boldsymbol{\beta})\mathbf{X}(\boldsymbol{\beta})$ , pourvu que ces deux matrices soient évaluées en un point suffisamment proche de  $\boldsymbol{\beta}_0$ . Nous avons besoin de cette condition parce que (6.10) n'est vérifiée qu'en  $\boldsymbol{\beta}_0$ , puisque ce résultat (6.10) dépend du fait que  $\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}_0) = \mathbf{u}_0$ ; consulter la discussion qui a précédé (5.38). La continuité de  $\mathbf{x}(\boldsymbol{\beta})$  suggère toutefois que (6.10) doit être approximativement vraie pour des valeurs de  $\boldsymbol{\beta}$  proches de  $\boldsymbol{\beta}_0$ .

<sup>1</sup> Si la matrice  $\mathbf{X}$  était normalisée de manière à ce que toutes ses colonnes aient la même longueur — par exemple une longueur unitaire — l'usage du nombre condition serait pertinent pour déceler la colinéarité de  $\mathbf{X}^\top\mathbf{X}$ . Voir Belsley, Kuh, et Welsch (1980, Chapitre 3).

Par conséquent, il serait extrêmement surprenant que la régression de Gauss-Newton ne réussisse pas à prévoir la colinéarité lorsque la matrice Hessienne est en réalité quasi singulière.

A titre d'exemple, considérons le modèle de régression non linéaire

$$y_t = \beta_1 + \beta_2 z_t^{\beta_3} + u_t. \quad (6.11)$$

Pour ce modèle, la  $t^{\text{ième}}$  ligne de  $\mathbf{X}(\boldsymbol{\beta})$  est

$$\begin{bmatrix} 1 & z_t^{\beta_3} & \beta_2 z_t^{\beta_3} \log(z_t) \end{bmatrix}. \quad (6.12)$$

A partir de là on voit immédiatement que la matrice  $\mathbf{X}^\top(\boldsymbol{\beta})\mathbf{X}(\boldsymbol{\beta})$  sera singulière si jamais  $\beta_2$ , ou  $\beta_3$ , est nul. Dans le premier cas, la troisième colonne de  $\mathbf{X}(\boldsymbol{\beta})$  sera composée de zéros, alors que dans le second, la deuxième colonne sera équivalente à la colonne correspondant à la constante. Ainsi ce modèle est asymptotiquement non identifié en des points tels que  $\beta_2 = 0$  ou  $\beta_3 = 0$ , bien qu'il soit identifié par presque tous les ensembles de données.

Si un modèle n'est pas identifié asymptotiquement même s'il est identifié par un ensemble de données précis, il est probable qu'il soit d'un intérêt restreint, puisque sans l'identification asymptotique, opérer des inférences correctes est impossible. De plus, si un modèle n'est pas identifié asymptotiquement en  $\boldsymbol{\beta}_0$ , les moindres carrés non linéaires ne produisent pas des estimations convergentes. D'un autre côté, un modèle qui est identifié asymptotiquement peut ne pas être identifié par un ensemble de données particulier. Il sera utile d'acquérir davantage de données en vue d'estimer un tel modèle.

Même pour des valeurs des paramètres pour lesquelles il est identifié, c'est-à-dire des valeurs différentes de  $\beta_2 = 0$  et  $\beta_3 = 0$ , le modèle (6.11) sera sûrement mal identifié. Il est clair d'après (6.12) que ce sera le cas le plus probable, en fonction des données et des valeurs des paramètres. La deuxième colonne de  $\mathbf{X}(\boldsymbol{\beta})$  est très similaire à la troisième colonne, chaque élément de la deuxième colonne multiplié par une constante et par  $\log(z_t)$  étant égal à l'élément correspondant de la troisième colonne. A moins que l'étendue de  $z_t$  ne soit très importante, ou qu'il y ait quelques valeurs de  $z_t$  très proches de zéro,  $z_t^{\beta_3}$  et  $\beta_2 \log(z_t) z_t^{\beta_3}$  tendront à être fortement corrélées. Par exemple, si  $\mathbf{z}^\top$  est composé des cinq observations [1 2 3 4 5], qui couvrent une étendue plus large que ne le feraient la plupart des régresseurs dans les applications économétriques, et si  $\beta_3$  est égal à l'unité, la corrélation entre les deux vecteurs s'établit à 0.9942; si  $\mathbf{z}^\top$  est composé des cinq observations [5 6 7 8 9], la corrélation atteint l'importante valeur de 0.9996. Remarquons que l'insuffisance d'identification du modèle dépend autant des données que des valeurs des paramètres et de la structure du modèle.

Nous avons déjà noté que la régression de Gauss-Newton

$$\mathbf{y} - \hat{\mathbf{x}} = \hat{\mathbf{X}}\mathbf{b} + \text{résidus},$$

où  $\mathbf{x}(\boldsymbol{\beta})$  et  $\mathbf{X}(\boldsymbol{\beta})$  sont évalués aux avec les estimations NLS  $\hat{\boldsymbol{\beta}}$ , produit une estimation valide de la matrice de covariance de  $\hat{\boldsymbol{\beta}}$  lorsque le DGP est un cas particulier du modèle qui a été estimé. Il est clair que la GNR

$$\mathbf{y} - \mathbf{x}_0 = \mathbf{X}_0 \mathbf{b} + \text{résidus}, \quad (6.13)$$

où  $\mathbf{x}(\boldsymbol{\beta})$  et  $\mathbf{X}(\boldsymbol{\beta})$  sont évaluées en  $\boldsymbol{\beta}_0$ , devrait également entraîner une estimation valide de  $\mathbf{V}(\hat{\boldsymbol{\beta}})$ , bien qu'en pratique cette régression ne soit pas envisageable parce qu'elle requiert la connaissance du DGP. Quoi qu'il en soit, il est utile de considérer  $\mathbf{V}(\hat{\boldsymbol{\beta}})$  comme ayant été générée par cette équation. Si l'on fait abstraction des finesses de l'analyse asymptotique, de façon à éviter la manipulation des divers facteurs de  $n$ , on peut utiliser simplement l'équation (6.13) pour mettre en lumière le problème de colinéarité dans les modèles de régression linéaire et dans les modèles de régression non linéaire.

Supposons que l'on porte quelque intérêt à la mesure de l'estimation NLS d'un élément unique de  $\boldsymbol{\beta}$ , que nous pouvons appeler  $\beta_1$  sans réduire la portée de l'analyse. Il est toujours possible de partitionner  $\mathbf{X}(\boldsymbol{\beta})$  en  $[\mathbf{x}_1(\boldsymbol{\beta}) \quad \mathbf{X}_2(\boldsymbol{\beta})]$ , où  $\mathbf{x}_1(\boldsymbol{\beta})$  désigne la colonne unique de  $\mathbf{X}(\boldsymbol{\beta})$  correspondant à  $\beta_1$ , et  $\mathbf{X}_2(\boldsymbol{\beta})$  désigne les  $k - 1$  colonnes restantes. Si l'on ne rapporte pas l'indice "0" par souci de clarté, la GNR (6.13) devient alors

$$\mathbf{u} = \mathbf{x}_1 b_1 + \mathbf{X}_2 \mathbf{b}_2 + \text{résidus}.$$

D'après le Théorème FWL, l'estimation de  $b_1$  pour cette régression artificielle sera numériquement identique à celle de la régression

$$\mathbf{M}_2 \mathbf{u} = \mathbf{M}_2 \mathbf{x}_1 b_1 + \text{résidus}, \quad (6.14)$$

où  $\mathbf{M}_2 \equiv \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1} \mathbf{X}_2^\top$  est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}_2)$ , le complémentaire orthogonal de l'espace engendré par  $\mathbf{X}_2$ .

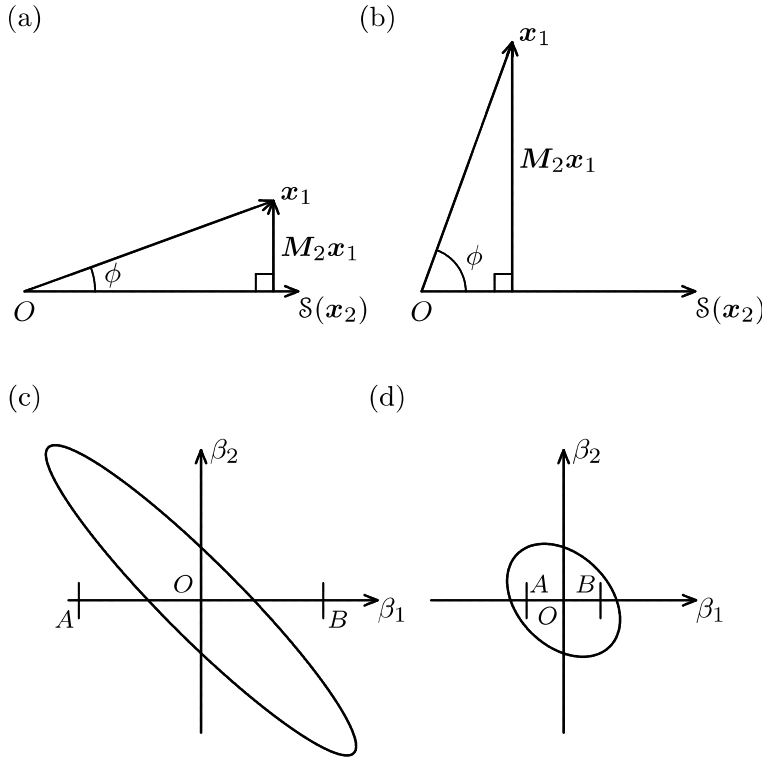
La régression (6.14) ne possède qu'un seul régresseur,  $\mathbf{M}_2 \mathbf{x}_1$ . Il est facile de voir que

$$V(\hat{b}_1) = \sigma_0^2 (\mathbf{x}_1^\top \mathbf{M}_2 \mathbf{x}_1)^{-1} = \frac{\sigma_0^2}{\mathbf{x}_1^\top \mathbf{M}_2 \mathbf{x}_1}.$$

C'est asymptotiquement la même que la variance de  $\hat{\beta}_1$ . Notons que  $\mathbf{x}_1^\top \mathbf{M}_2 \mathbf{x}_1$  est tout simplement la somme des résidus au carré de la régression

$$\mathbf{x}_1 = \mathbf{X}_2 \mathbf{c}_2 + \text{résidus}. \quad (6.15)$$

Ainsi, on en conclut que la variance de  $\hat{\beta}_1$  est proportionnelle à l'inverse de la somme des carrés des résidus de la régression (6.15). Lorsque  $\mathbf{x}_1$  est bien expliqué par les autres colonnes de  $\mathbf{X}$ , cette SSR sera faible et la variance de  $\hat{\beta}_1$  sera par conséquent importante. Lorsque ce n'est pas le cas et que  $\mathbf{x}_1$  est assez mal expliqué par les autres colonnes de  $\mathbf{X}$ , cette SSR sera forte, et la variance



**Figure 6.1** Colinéarité et précision de l'estimation

de  $\hat{\beta}_1$  deviendra faible. Ces deux cas ont été illustrés dans la Figure 6.1, pour le cas où il n'y a que deux régresseurs. Il faudrait comparer cette figure à la Figure 3.3, car pour l'essentiel les mêmes résultats apparaissent dans les deux figures, puisque la longueur de l'intervalle de confiance d'un paramètre donné est proportionnel à la racine carrée de la variance estimée de ce paramètre.

Le régresseur  $\mathbf{x}_2$ , qui représente sur la figure tous les régresseurs autres que le régresseur  $\mathbf{x}_1$  qui nous intéresse, est le même de chaque côté de la figure. D'autre part, le régresseur  $\mathbf{x}_1$  est orienté différemment par rapport à  $\mathbf{x}_2$  sur les deux parties. Pour simplifier les deux régresseurs sont de même longueur et la seule variable est l'angle  $\phi$ , qu'ils forment ensemble. Dans le schéma (a) en haut à gauche,  $\mathbf{x}_2$  explique  $\mathbf{x}_1$  relativement bien, alors l'angle  $\phi$  et la somme des résidus au carré de (6.15) sont relativement faibles. Cette dernière est le carré de la longueur du vecteur de résidus  $\mathbf{M}_2\mathbf{x}_1$ . Dans le schéma (b),  $\mathbf{x}_2$  explique beaucoup moins bien  $\mathbf{x}_1$ , de sorte que l'angle  $\phi$  et la SSR de (6.15) sont assez importants. Clairement le degré de colinéarité entre  $\mathbf{x}_1$  et  $\mathbf{x}_2$  est plus fort lorsque  $\phi$  est faible, ou de façon équivalente, lorsque  $\mathbf{x}_2$  explique bien  $\mathbf{x}_1$ .

Dans les schémas (c) et (d), nous avons représenté les ellipses et intervalles de confiance pour  $\hat{\beta}_1$  correspondant aux régresseurs illustrés dans les schémas (a) et (b). Pour simplifier, nous les avons représentés dans le cas  $\hat{\boldsymbol{\beta}} = \mathbf{0}$ :

dans tout exemple réel le seul aspect différent des figures serait l'origine qui se déplacerait n'importe où, laissant inchangées les tailles et les positions relatives des autres éléments. Remarquons que les deux schémas du bas ont été dessinés à la même échelle. Les intervalles de confiance, qui sont les segments  $AB$  dans la figure, sont donc de longueurs très différentes. Comme nous l'avons expliqué à la Section 3.3, un plus haut degré de colinéarité fait apparaître une région de confiance elliptique de plus grande excentricité, et cela se voit aisément sur la figure. Pour la situation dépeinte, dans laquelle seul l'angle  $\phi$  varie, il peut être montré que l'aire de l'ellipse de confiance et la longueur de l'intervalle de confiance de  $\hat{\beta}_1$  sont inversement proportionnelles à  $\sin \phi$ . Par conséquent, si  $\phi$  tend vers zéro, et si les deux régresseurs se rapprochent en devenant colinéaires, la longueur de l'intervalle de confiance tend vers l'infini. Cela, bien évidemment, ne fait que refléter le fait que la variance de  $\hat{\beta}_1$  tend vers l'infini lorsque  $\phi$  tend vers zéro.

Le phénomène de colinéarité se manifeste lorsqu'une ou plusieurs colonnes de  $\mathbf{X}$  sont extrêmement bien expliquées par les colonnes restantes, et lorsque les estimations des paramètres associés à ces colonnes sont très imprécises. Un moyen très simple de caractériser la présence ou l'absence de colinéarité, étant donné qu'elle affecte l'estimation du paramètre unique  $\beta_1$ , est de considérer le rapport de  $\mathbf{x}_1^\top \mathbf{M}_2 \mathbf{x}_1$  par  $\mathbf{x}_1^\top \mathbf{M}_\iota \mathbf{x}_1$ , où  $\mathbf{M}_\iota \equiv \mathbf{I} - \iota(\iota^\top \iota)^{-1} \iota^\top$  est la matrice qui opère la déviation par rapport à la moyenne (comme d'habitude,  $\iota$  désigne un vecteur composé de  $n$  uns). Le numérateur de ce rapport mesure la variation de  $\mathbf{x}_1$  qui n'est pas expliquée par la variation de  $\mathbf{X}_2$ , tandis que le dénominateur mesure la variation de  $\mathbf{x}_1$  autour de sa moyenne. Si ce rapport est très faible, la colinéarité peut se révéler être un problème.

Lorsqu'un modèle est insuffisamment identifié pour une quelconque valeur de  $\beta$ , disons  $\beta^*$ , la régression de Gauss-Newton manifestera de façon systématique un niveau important de colinéarité lorsqu'elle est évaluée en  $\beta^*$ . Si l'on éprouve quelques difficultés à obtenir des estimations NLS, ou si l'on n'a pas encore débuté l'estimation d'un modèle et que l'on s'attend à rencontrer quelques obstacles, il peut devenir très utile de savoir si, oui ou non, la régression de Gauss-Newton est en proie à un phénomène conséquent de colinéarité pour des valeurs plausibles des paramètres. Si c'est effectivement le cas, alors le modèle est sûrement insuffisamment identifié par les données, et il peut donc s'avérer impossible d'obtenir des estimations d'une précision satisfaisante de ce modèle avec cet ensemble de données, ou même de localiser le minimum de  $SSR(\beta)$  (voir Section 6.8).

Que faire lorsque l'on est confronté à un modèle de régression non linéaire qui est insuffisamment identifié? Il y a fondamentalement deux options qui se présentent: recueillir davantage de données, ou estimer un modèle plus simple, peut-être le modèle originel après lui avoir imposé quelques contraintes. S'il s'avère irréalisable de recueillir davantage de données, alors il faut accepter l'idée selon laquelle les données dont on dispose contiennent une quantité limitée d'informations, et adapter le modèle à la situation. L'estimation de

modèles qui se révèlent trop complexes est une des erreurs les plus fréquentes chez les économètres inexpérimentés. Un modèle tel que (6.11), par exemple, requiert un nombre important de données, et sera sans doute très difficile à estimer avec de nombreux ensembles de données. Il ne sera possible d'obtenir des estimations précises de  $\beta_2$  et  $\beta_3$  pour ce genre de modèle que si le nombre d'observations est très important et/ou le champ de  $z_t$  est très étendu.

## 6.4 TESTS DE CONTRAINTES

L'usage le plus connu de la régression de Gauss-Newton est celui qui consiste à offrir un moyen simple de calcul de statistiques de test. Une fois obtenues les estimations contraintes, une variante de la GNR peut être utilisée pour les tests de tous types de contraintes d'égalité sur  $\beta$  sans avoir à estimer le modèle non contraint. Ces tests sont basés sur le principe du multiplicateur de Lagrange dont nous avons discuté à la Section 3.6. De nombreuses statistiques de test, numériquement différentes mais asymptotiquement équivalentes, peuvent se calculer en se basant sur la GNR. Dans ce chapitre nous ne traitons que les statistiques de test qui sont basées sur des estimations NLS contraintes. Comme nous le constaterons à la Section 6.7, la GNR peut aussi être employée dans le calcul de tests basés sur toute estimation convergente au taux  $n^{1/2}$ .

Ecrivons l'hypothèse nulle et l'hypothèse alternative sous la forme

$$\left. \begin{array}{l} H_0 : \mathbf{y} = \mathbf{x}(\beta_1, \mathbf{0}) + \mathbf{u} \\ H_1 : \mathbf{y} = \mathbf{x}(\beta_1, \beta_2) + \mathbf{u} \end{array} \right\} \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}),$$

où  $\beta_1$  est de dimension  $(k - r) \times 1$ , et  $\beta_2$  est de dimension  $r \times 1$ . En ne prenant en compte que des restrictions de nullité, nous ne limitons nullement la généralité des résultats, car, comme nous en avons déjà discuté, tout ensemble de  $r$  contraintes d'égalité peut être converti en un ensemble de  $r$  contraintes de nullité, grâce à une reparamétrisation adéquate. Nous supposerons que le modèle non contraint est identifié asymptotiquement au voisinage du DGP, qui est supposé appartenir à la famille des DGP

$$\mathbf{y} = \mathbf{x}(\beta_{01}, \mathbf{0}) + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I})$$

de sorte que c'est  $H_0$  qui a généré réellement les données. La matrice  $\mathbf{X}(\beta)$  peut être partitionnée en deux,  $\mathbf{X}_1(\beta)$  et  $\mathbf{X}_2(\beta)$ , correspondant à  $\beta_1$  et  $\beta_2$ , et de dimensions respectives  $n \times (k - r)$  et  $n \times r$ .

Il n'est pas évident que toutes les contraintes d'égalité puissent être converties en des contraintes de nullité à l'aide d'une reparamétrisation adéquate. Dans la Section 3.6 par exemple, nous considérons les contraintes linéaires du type  $\mathbf{R}\gamma = \mathbf{r}$ , où  $\mathbf{R}$  est une matrice de dimension  $r \times k$  et  $\mathbf{r}$  est un vecteur à  $r$  composantes. Ici  $\gamma$  désigne le vecteur à  $k$  composantes sur lequel portent



les contraintes; nous ferons usage de  $\beta$  pour désigner une paramétrisation alternative pour laquelle toutes les contraintes sont des contraintes de nullité. Pour cette paramétrisation la matrice  $\mathbf{R}$  prend la forme  $[\mathbf{0} \quad \mathbf{I}_r]$ , et le vecteur  $\mathbf{r}$  est composé de zéros. Ainsi il apparaît que

$$\mathbf{R}\gamma - \mathbf{r} = \mathbf{0} = [\mathbf{0} \quad \mathbf{I}_r]\beta, \quad (6.16)$$

où  $\gamma$  et  $\beta$  désignent les vecteurs de paramètres dans deux paramétrisations différentes. Evidemment on peut associer  $\beta_i$  à  $\gamma_i$  pour  $i = 1, \dots, k - r$ , alors que dans l'équation (6.16) il apparaît que

$$\beta_i = \sum_{j=1}^k R_{ij}\gamma_j - r_i$$

pour  $i = k - r + 1, \dots, k$ . Ainsi la conversion des contraintes linéaires  $\mathbf{R}\gamma = \mathbf{r}$  paramétrisées en  $\gamma$  en des contraintes de nullité paramétrisées en  $\beta$  est directe. Il est aussi possible de convertir des contraintes non linéaires, telles que celles dont nous avons parlé à la Section 5.7, en des contraintes de nullité. Il faut simplement définir les nouveaux paramètres en fonction des anciens paramètres. Par exemple, si une des contraintes était  $\gamma_1^2 - 5 = 0$ , on pourrait définir un nouveau paramètre  $\beta_1$  de façon à le rendre égal à  $\gamma_1^2 - 5$ . Bien sûr, ce type de reparamétrisation non linéaire n'est pas toujours aisé dans la pratique s'il y a plusieurs contraintes non linéaires compliquées. Par chance, il n'est pas nécessaire de reparamétriser le modèle de façon à n'obtenir que des contraintes de nullité en vue d'utiliser les résultats théoriques basés sur l'hypothèse que ces contraintes sont de ce genre.

L'identification asymptotique implique que la matrice  $n^{-1}\mathbf{X}_0^\top \mathbf{X}_0$  doit tendre vers une matrice définie positive. Cette hypothèse élimine un certain nombre de modèles et de contraintes. Par exemple, il ne nous serait pas possible de traiter un modèle tel que

$$y_t = \beta_1 + \beta_2 \exp(\beta_3 z_t) + u_t$$

soumis aux contraintes  $\beta_2 = 0$  ou  $\beta_3 = 0$ , puisqu'alors le modèle ne serait pas identifiable asymptotiquement. Nous noterons l'estimation de  $\beta$  sous  $H_0$  par  $\tilde{\beta}$ , pour la distinguer du vecteur d'estimations non contraintes  $\hat{\beta}$ ; toutes les quantités repérées par un  $\sim$  sont évaluées par les estimations contraintes. Dans ce cas,  $\tilde{\beta} \equiv [\tilde{\beta}_1 : \mathbf{0}]$ .

A présent nous allons examiner la distribution de quelques statistiques de test étroitement liées, qui s'obtiendraient en exécutant une régression de Gauss-Newton avec  $\beta$  évalué en  $\tilde{\beta}$ . Toutes ces statistiques de test sont asymptotiquement équivalentes à la statistique de test LM (3.47), ou sa variante, la forme du score (3.48), qui a été dévoilée à la Section 5.7 comme étant asymptotiquement distribuée suivant la  $\chi^2(r)$  sous l'hypothèse nulle. La régression de Gauss-Newton évaluée en  $\beta = \tilde{\beta}$  est

$$\mathbf{y} - \tilde{\mathbf{x}} = \tilde{\mathbf{X}}_1 \mathbf{b}_1 + \tilde{\mathbf{X}}_2 \mathbf{b}_2 + \text{résidus}. \quad (6.17)$$

Ce qui ressemble énormément à la régression (3.49), que nous avons introduite à la Section 3.6 en tant que moyen de calcul de la statistique du multiplicateur de Lagrange. La seule différence entre les deux est que la régressande n'a pas été divisée par une estimation de  $\sigma$ ; comme nous allons le voir bientôt, la division de la régressande par une estimation de  $\sigma$  est facultative.

Par l'intermédiaire du Théorème FWL, nous voyons que la régression (6.17) entraîne exactement les mêmes estimations de  $\mathbf{b}_2$ , c'est-à-dire  $\tilde{\mathbf{b}}_2$ , et exactement la même somme des résidus au carré que la régression

$$\mathbf{y} - \tilde{\mathbf{x}} = \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2 \mathbf{b}_2 + \text{résidus}, \quad (6.18)$$

où  $\tilde{\mathbf{M}}_1$  est la matrice qui projette sur  $\mathcal{S}^\perp(\tilde{\mathbf{X}}_1)$ . La régressande n'est pas ici multipliée par  $\tilde{\mathbf{M}}_1$  parce que les conditions du premier ordre impliquent que  $\mathbf{y} - \tilde{\mathbf{x}}$  se situe déjà dans  $\mathcal{S}^\perp(\tilde{\mathbf{X}}_1)$ , de sorte que  $\tilde{\mathbf{M}}_1(\mathbf{y} - \tilde{\mathbf{x}}) = (\mathbf{y} - \tilde{\mathbf{x}})$ . La somme des résidus au carré de la régression (6.18) est

$$(\mathbf{y} - \tilde{\mathbf{x}})^\top (\mathbf{y} - \tilde{\mathbf{x}}) - (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}}_2 (\tilde{\mathbf{X}}_2^\top \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2^\top (\mathbf{y} - \tilde{\mathbf{x}}).$$

Puisque  $\mathbf{y} - \tilde{\mathbf{x}}$  se trouve dans  $\mathcal{S}^\perp(\tilde{\mathbf{X}}_1)$ , il doit être orthogonal à  $\tilde{\mathbf{X}}_1$ . Ainsi si nous n'avons pas compris  $\tilde{\mathbf{X}}_2$  dans la régression, la SSR aurait été égale à  $(\mathbf{y} - \tilde{\mathbf{x}})^\top (\mathbf{y} - \tilde{\mathbf{x}})$ . Par conséquent la réduction dans la SSR de la régression (6.17) consécutive à l'addition de  $\tilde{\mathbf{X}}_2$  est

$$(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}}_2 (\tilde{\mathbf{X}}_2^\top \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2^\top (\mathbf{y} - \tilde{\mathbf{x}}). \quad (6.19)$$

Cette quantité correspond également à la somme des carrés expliqués (autour de zéro) de la régression (6.17), une fois de plus à cause du pouvoir explicatif de  $\tilde{\mathbf{X}}_1$ . Comme nous le montrons maintenant, cette quantité divisée par n'importe quelle estimation convergente de  $\sigma^2$ , est distribuée asymptotiquement selon une  $\chi^2(r)$  sous l'hypothèse nulle.

Pour s'en rendre compte, observons que

$$n^{-1/2}(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}}_2 \stackrel{a}{=} n^{-1/2} \mathbf{u}^\top \mathbf{M}_1 \mathbf{X}_2 \equiv \boldsymbol{\nu}^\top,$$

où  $\mathbf{M}_1 \equiv \mathbf{M}_1(\boldsymbol{\beta}_0)$  et  $\mathbf{X}_2 \equiv \mathbf{X}_2(\boldsymbol{\beta}_0)$ . L'égalité asymptotique provient ici du fait que  $\tilde{\mathbf{u}} \stackrel{a}{=} \mathbf{M}_1 \mathbf{u}$ , qui est le résultat (6.09) dans le cas où le modèle est estimé sous la contrainte  $\boldsymbol{\beta}_2 = \mathbf{0}$ . La matrice de covariance du vecteur aléatoire  $\boldsymbol{\nu}$  de dimension  $r$  est

$$\begin{aligned} E(\boldsymbol{\nu} \boldsymbol{\nu}^\top) &= E(n^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{u} \mathbf{u}^\top \mathbf{M}_1 \mathbf{X}_2) = n^{-1} \mathbf{X}_2^\top \mathbf{M}_1 (\sigma_0^2 \mathbf{I}) \mathbf{M}_1 \mathbf{X}_2 \\ &= n^{-1} \sigma_0^2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2) \equiv \sigma_0^2 \mathbf{V}. \end{aligned}$$

La convergence de  $\tilde{\boldsymbol{\beta}}$  et les conditions de régularité pour le Théorème 5.1 impliquent que

$$n^{-1} \tilde{\mathbf{X}}_2^\top \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2 \stackrel{a}{=} n^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2 = \mathbf{V}.$$

Sous les conditions de régularité du Théorème 5.2, il nous est possible d'appliquer un théorème de la limite centrale et de conclure que  $\boldsymbol{\nu}$  est distribué asymptotiquement selon une normale centrée et de matrice de covariance  $\sigma_0^2 \mathbf{V}$ . Ainsi toute statistique de test construite en divisant (6.19) par une estimation consistante de  $\sigma^2$  est asymptotiquement égale à

$$\frac{1}{\sigma_0^2} \boldsymbol{\nu}^\top \mathbf{V}^{-1} \boldsymbol{\nu}.$$

Nous voyons immédiatement que cette quantité est distribuée asymptotiquement suivant une  $\chi^2(r)$ , puisque c'est une forme quadratique en un vecteur aléatoire à  $r$  composantes qui est asymptotiquement d'espérance nulle et d'une matrice de covariance  $\sigma_0^2 \mathbf{V}$ ; voir Annexe B.

Parce que nous voulons utiliser n'importe quelle estimation convergente de  $\sigma^2$ , ce résultat justifie l'usage de plusieurs statistiques de test différentes. Une possibilité est d'utiliser la somme des carrés expliqués de la régression (3.49) dans laquelle la régressande de (6.17) a été divisée par une estimation convergente de  $\sigma$ . Cependant les deux statistiques de test les plus courantes sont  $n$  fois le  $R^2$  non centré de la régression (6.17) et le Fisher habituel pour  $\mathbf{b}_2 = \mathbf{0}$  de la même régression. Pour constater que la forme  $nR^2$  du test est correcte, observons que puisque  $R_u^2$ , le  $R^2$  non centré, est égal à la somme des carrés expliqués divisée par la somme des carrés totaux,

$$\begin{aligned} nR_u^2 &= \frac{n(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}}_2 (\tilde{\mathbf{X}}_2^\top \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2^\top (\mathbf{y} - \tilde{\mathbf{x}})}{(\mathbf{y} - \tilde{\mathbf{x}})^\top (\mathbf{y} - \tilde{\mathbf{x}})} \\ &= \frac{\|\mathbf{P}_{\tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2}(\mathbf{y} - \tilde{\mathbf{x}})\|^2}{\|\mathbf{y} - \tilde{\mathbf{x}}\|^2/n}. \end{aligned}$$

Cette variante du test emploie implicitement  $\tilde{\sigma}^2$ , l'estimation contrainte du maximum de vraisemblance pour estimer  $\sigma^2$ . Comme nous l'avons remarqué, cette estimation tendra à être trop faible, du moins lorsque l'hypothèse nulle est exacte. Il serait sûrement plus prudent d'utiliser  $(n - k + r)R_u^2$  comme statistique de test, puisqu'elle utiliserait implicitement  $\tilde{s}^2$ , l'estimation OLS de  $\sigma^2$  dans le modèle contraint, au lieu de  $\tilde{\sigma}^2$ ; la statistique de test qui en résulterait serait égale à la ESS de (3.49).

Un problème d'ordre pratique qui se pose avec des tests basés sur le  $R_u^2$  de la régression artificielle est que de nombreux progiciels de régression n'affichent pas le  $R^2$  non centré. Dans la plupart des cas, cela ne posera pas de difficulté parce que le  $R_u^2$  sera identique au  $R^2$  centré ordinaire. Cela sera vérifié chaque fois que le modèle contraint  $\mathbf{x}(\boldsymbol{\beta}_1, \mathbf{0})$  comprend l'équivalent d'un terme constant de sorte que  $\mathbf{y} - \tilde{\mathbf{x}}$  sera de moyenne nulle. Dans le cas où  $\mathbf{y} - \tilde{\mathbf{x}}$  n'a pas une moyenne nulle, le  $R^2$  non centré différera cependant du  $R^2$  centré, et ce dernier ne produit pas une statistique de test correcte. Il est crucial que les utilisateurs de statistiques de test  $nR^2$  et  $(n - k + r)R^2$  soient

avertis de cette éventualité, et vérifient que la régressande pour la régression de test (6.17) est bien de moyenne nulle.

Il est également envisageable, et sûrement préférable, d'utiliser le Fisher habituel pour  $\mathbf{b}_2 = \mathbf{0}$  dans la régression de Gauss-Newton. Si RSSR et USSR désignent les sommes des résidus au carré contraints et non contraints de la régression (6.17), ce Fisher est

$$\begin{aligned} & \frac{(\text{RSSR} - \text{USSR})/r}{\text{USSR}/(n-k)} \\ &= \frac{(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}}_2 (\tilde{\mathbf{X}}_2^\top \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2^\top (\mathbf{y} - \tilde{\mathbf{x}})/r}{((\mathbf{y} - \tilde{\mathbf{x}})^\top (\mathbf{y} - \tilde{\mathbf{x}}) - (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{X}}_2 (\tilde{\mathbf{X}}_2^\top \tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2^\top (\mathbf{y} - \tilde{\mathbf{x}}))/(n-k)}. \end{aligned} \quad (6.20)$$

Le dénominateur est ici l'estimation OLS de  $\sigma^2$  de (6.17), qui tend vers  $\sigma_0^2$  lorsque  $n \rightarrow \infty$  sous  $H_0$ . Le numérateur est  $1/r$  fois l'expression (6.19). Ainsi, il est clair que  $r$  fois (6.20) sera distribuée asymptotiquement selon une  $\chi^2(r)$ .

Avec des échantillons finis, la comparaison entre (6.20) et la distribution de  $F(r, n-k)$  est aussi correcte que la comparaison entre  $r$  fois (6.20) et la distribution du  $\chi^2(r)$ . En fait, il est évident que le Fisher (6.20) possède de meilleures propriétés en échantillons finis que la statistique  $nR^2$  basée sur la même régression de Gauss-Newton; consulter Kiviet (1986). Cette évidence va dans le même sens que la théorie, parce que, comme nous l'avons vu,  $\hat{\mathbf{u}} \stackrel{a}{=} \mathbf{M}_0 \mathbf{u}$ . Aussi l'usage de la distribution  $F$ , qui considère l'estimation  $s^2$  basée sur les résidus NLS comme si elle était basée sur les résidus OLS, est plus pertinent que l'usage de la distribution  $\chi^2$ , qui traite  $\hat{\sigma}^2$  comme étant basée sur les aléas au lieu des résidus. Sur la base de la théorie et de l'évidence d'une part, et de la commodité de l'usage d'un test qui aurait la même forme pour les régressions de Gauss-Newton comme pour d'autres régressions plus authentiques d'autre part, nous conseillons donc l'usage d'un test en  $F$  plutôt qu'un test en  $nR^2$  ou qu'un test numériquement comparable basé sur la régression (3.49).

L'expression (6.20) peut se simplifier en notant que c'est tout simplement  $(n-k)/r$  fois le ratio des longueurs au carré de deux vecteurs

$$\frac{n-k}{r} \times \frac{\|\mathbf{P}_{\tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2}(\mathbf{y} - \tilde{\mathbf{x}})\|^2}{\|\mathbf{M}_{\tilde{\mathbf{X}}}(\mathbf{y} - \tilde{\mathbf{x}})\|^2}. \quad (6.21)$$

Le numérateur est la longueur au carré du vecteur  $\mathbf{P}_{\tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2}(\mathbf{y} - \tilde{\mathbf{x}})$ , qui est le vecteur de résidus  $\mathbf{y} - \tilde{\mathbf{x}}$  projeté sur  $\mathcal{S}(\tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2)$ . Le dénominateur est la longueur au carré du vecteur  $\mathbf{M}_{\tilde{\mathbf{X}}}(\mathbf{y} - \tilde{\mathbf{x}}) = \mathbf{M}_{\tilde{\mathbf{M}}_1 \tilde{\mathbf{X}}_2}(\mathbf{y} - \tilde{\mathbf{x}})$ , qui est le vecteur de résidus projeté en dehors de  $\mathcal{S}(\tilde{\mathbf{X}}) = \mathcal{S}(\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ . La géométrie du test en  $F$  dans ce cas est identique à la géométrie du test en  $F$  dans un cas de régression linéaire, dont nous avons déjà discuté à la Section 3.5. La seule différence est que  $\tilde{\mathbf{X}}_1$  et  $\tilde{\mathbf{X}}_2$  sont des fonctions de l'estimation contrainte  $\tilde{\boldsymbol{\beta}}$ .

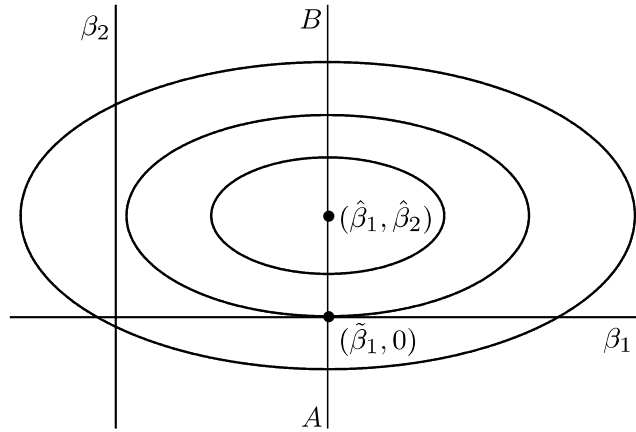
Nous avons démontré à la Section 3.5 que, pour les modèles de régression linéaire, le Student portant sur l'hypothèse que la valeur d'un unique paramètre est nulle est numériquement égal à la racine carrée d'un Fisher portant sur la même hypothèse nulle. Puisqu'il s'agit d'un résultat numérique, il est valable pour les régressions artificielles autant que pour les véritables régressions. Alors lorsque  $b_2$  est un scalaire, le Student portant sur  $\tilde{b}_2$  à partir de la GNR (6.17) est aussi valable que toutes les autres statistiques de test dont nous avons déjà parlé.

Pourquoi la régression des résidus du modèle contraint sur les dérivées de  $\mathbf{x}(\boldsymbol{\beta})$  nous permet-elle de calculer des statistiques de test valides? Pourquoi avons-nous besoin de comprendre toutes les dérivées, et pas seulement celles correspondant aux paramètres qui sont contraints? La discussion que nous avons tenue plus haut fournit des réponses formelles à ces questions, mais peut-être pas celles qui sont intuitivement plaisantes. Considérons donc l'importance d'un point de vue légèrement différent. A la Section 5.7, nous montrions que les statistiques de Wald, LR et LM pour tester le même ensemble de contraintes sont asymptotiquement égales à la même variable aléatoire sous l'hypothèse nulle, et que cette variable aléatoire est asymptotiquement distribuée selon une  $\chi^2(r)$ . Pour les modèles de régression non linéaire que nous avons traités, la statistique LR est simplement la différence entre  $SSR(\hat{\boldsymbol{\beta}})$  et  $SSR(\tilde{\boldsymbol{\beta}})$ , divisée par n'importe quelle estimation convergente de  $\sigma^2$ . Pour voir pourquoi la statistique LM est valide et pourquoi la GNR doit inclure les dérivées par rapport à tous les paramètres, nous allons envisager la statistique LM basée sur la GNR comme une approximation quadratique de cette statistique LR. Et cela a un sens parce que la GNR est elle-même une approximation linéaire du modèle de régression non linéaire.

Une façon de considérer la régression de Gauss-Newton est d'imaginer que c'est un moyen d'approximer la fonction  $SSR(\boldsymbol{\beta})$  par une fonction quadratique qui possède les mêmes dérivées premières et secondes au point  $\tilde{\boldsymbol{\beta}}$ . Cette fonction quadratique approximée, que nous nommerons  $SSR^*(\tilde{\boldsymbol{\beta}}, \mathbf{b})$ , est tout simplement la fonction somme des carrés de la régression artificielle. Elle se définit comme

$$SSR^*(\tilde{\boldsymbol{\beta}}, \mathbf{b}) = (\mathbf{y} - \tilde{\mathbf{x}} - \tilde{\mathbf{X}}\mathbf{b})^\top (\mathbf{y} - \tilde{\mathbf{x}} - \tilde{\mathbf{X}}\mathbf{b}).$$

La somme des carrés expliqués de la GNR correspond précisément à la différence entre  $SSR(\tilde{\boldsymbol{\beta}})$  et  $SSR^*(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{b}})$ . Si  $\tilde{\boldsymbol{\beta}}$  est relativement proche de  $\hat{\boldsymbol{\beta}}$ ,  $SSR^*(\cdot)$  doit fournir une bonne approximation de  $SSR(\cdot)$  au voisinage de  $\hat{\boldsymbol{\beta}}$ . En fait, pourvu que les contraintes soient exactes et la taille de l'échantillon suffisamment importante,  $\tilde{\boldsymbol{\beta}}$  et  $\hat{\boldsymbol{\beta}}$  seront proches l'un de l'autre parce qu'ils sont tous les deux convergents pour  $\boldsymbol{\beta}_0$ , de sorte que  $SSR^*(\cdot)$  doit fournir une bonne approximation de  $SSR(\cdot)$ . Ainsi  $SSR^*(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{b}})$  sera proche de  $SSR(\hat{\boldsymbol{\beta}})$  et la somme des carrés expliqués de la GNR sera par conséquent une bonne approximation de  $SSR(\tilde{\boldsymbol{\beta}}) - SSR(\hat{\boldsymbol{\beta}})$ . Lorsque nous divisons la somme des carrés expliqués par une estimation convergente de  $\sigma^2$ , la statistique de test LM qui en découle doit donc être similaire à la statistique de LR.



**Figure 6.2** Cas où la minimisation de  $SSR^{**}$  est satisfaisante

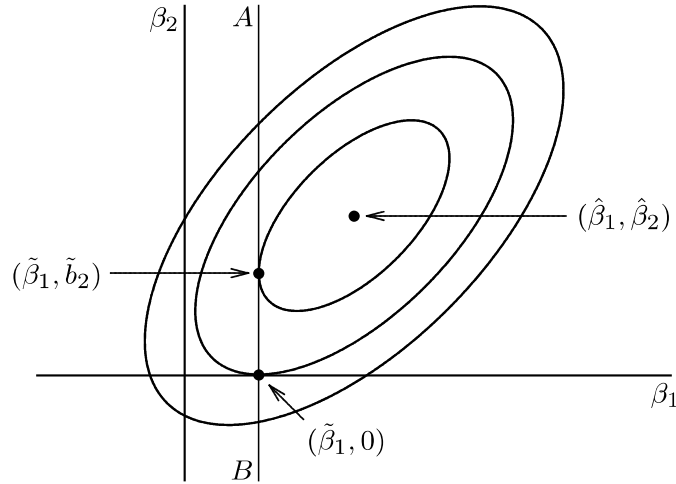
Il doit être clair désormais que la GNR doit inclure autant  $\tilde{\mathbf{X}}_1$  que  $\tilde{\mathbf{X}}_2$ . Si ce n'était pas le cas, la GNR ne minimiserait pas  $SSR^*(\tilde{\boldsymbol{\beta}}, \mathbf{b})$ , mais plutôt une autre approximation de  $SSR(\boldsymbol{\beta})$ ,

$$SSR^{**}(\tilde{\boldsymbol{\beta}}, \mathbf{b}_2) = (\mathbf{y} - \tilde{\mathbf{x}} - \tilde{\mathbf{X}}_2 \mathbf{b}_2)^\top (\mathbf{y} - \tilde{\mathbf{x}} - \tilde{\mathbf{X}}_2 \mathbf{b}_2).$$

Bien que  $SSR^*(\cdot)$  offre normalement une approximation raisonnablement satisfaisante de  $SSR(\cdot)$ ,  $SSR^{**}(\cdot)$  n'en fera pas de même, puisqu'elle ne dispose pas d'un nombre suffisant de paramètres libres. Lorsque l'on minimise  $SSR^{**}(\cdot)$ , cela n'est possible que dans les directions qui correspondent à  $\beta_2$  mais pas dans les directions qui correspondent à  $\beta_1$ .  $SSR^{**}(\cdot)$  ne sera satisfaisante que dans le cas où les contours de  $SSR(\cdot)$  forment approximativement une ellipse dont les axes sont sensiblement parallèles aux axes de  $\beta_1$  et  $\beta_2$ , de sorte que  $\hat{\beta}_1$  et  $\tilde{\beta}_1$  soient très proches. Dans les autres cas de figure, le minimum de  $SSR^*(\cdot)$  peut être très différent du minimum de  $SSR^{**}(\cdot)$ , et les valeurs correspondant au minimum de chaque fonction peuvent aussi être différentes.

Ceci est illustré sur les Figures 6.2 et 6.3 dans le cas où  $k = 2$  et  $r = 1$ . Dans la première figure, les vecteurs  $\tilde{\mathbf{x}}_1$  et  $\tilde{\mathbf{x}}_2$  sont orthogonaux, et les axes de l'ellipse formée par les contours de  $SSR(\cdot)$  sont précisément parallèles aux axes des abscisses et des ordonnées. Dans cette situation le minimum de  $SSR^{**}(\tilde{\beta}_1, \tilde{\mathbf{b}}_2)$ , qui doit se situer sur la ligne  $AB$ , coïncide avec le minimum de  $SSR(\beta_1, \beta_2)$  au point  $(\hat{\beta}_1, \hat{\beta}_2)$ . Dans la Figure 6.3  $\tilde{\mathbf{x}}_1$  et  $\tilde{\mathbf{x}}_2$  sont corrélés négativement de sorte que les axes de l'ellipse formée par les contours de  $SSR(\cdot)$  sont inclinés vers le haut à droite. Dans cette situation le minimum de  $SSR^{**}(\tilde{\beta}_1, \mathbf{b}_2)$ , qui une fois de plus doit se situer sur la ligne  $AB$ , est à l'évidence très différent du minimum de  $SSR(\beta_1, \beta_2)$ .

A l'évidence, le minimum de  $SSR^{**}(\cdot)$  ne peut être inférieur, et sera normalement supérieur au minimum de  $SSR^*(\cdot)$ . Ainsi, si l'on omettait par inadvertance  $\tilde{\mathbf{X}}_1$  dans la GNR et si l'on calculait  $nR_u^2$  ou une des autres valeurs de la statistique de test équivalente, nous obtiendrions une statistique de



**Figure 6.3** Cas où la minimisation de  $SSR^{**}$  est peu satisfaisante

test inférieure à la valeur correcte. C'est un phénomène utile dont il faut se souvenir. Dans certains cas, il peut s'avérer aisé de construire  $\tilde{\mathbf{X}}_2$  mais au contraire difficile de construire  $\tilde{\mathbf{X}}_1$ . Dans de telles situations il serait utile, en premier lieu, de régresser  $\tilde{\mathbf{u}}$  sur  $\tilde{\mathbf{X}}_2$  isolément. Si cette régression artificielle apporte une évidence définitivement défavorable envers l'hypothèse nulle, alors nous pourrions sans aucun doute décider de rejeter  $H_0$  sans avoir à exécuter la GNR correcte. Cependant, si une régression de  $\tilde{\mathbf{u}}$  sur  $\tilde{\mathbf{X}}_2$  isolément ne permet pas de rejeter l'hypothèse nulle, il n'est pas possible de tirer des conclusions sûres à partir de cette régression et il est nécessaire d'exécuter la vraie GNR pour pouvoir aboutir à des conclusions sûres.

La GNR (6.17) resterait correcte si l'on remplaçait  $\tilde{\mathbf{X}}_2$  par *n'importe quelle* matrice de dimension  $n \times r$ , disons  $\mathbf{Z}(\boldsymbol{\beta})$ , évaluée en  $\tilde{\boldsymbol{\beta}}$ , qui serait asymptotiquement non corrélée à  $\mathbf{u}$  sous l'hypothèse nulle, qui satisferait les mêmes conditions que  $\mathbf{X}(\boldsymbol{\beta})$ , et qui pourrait dépendre effectivement ou pas de  $\boldsymbol{\beta}$ . Ainsi, si le modèle qui doit être testé était

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (6.22)$$

et si  $\mathbf{Z}(\boldsymbol{\beta})$  était une matrice construite de sorte que, si le modèle était exact,  $n^{-1} \mathbf{Z}_0^\top \mathbf{u}$  tendrait vers un vecteur nul, il nous serait toujours possible d'effectuer la GNR

$$\tilde{\mathbf{u}} = \tilde{\mathbf{X}}\mathbf{b} + \tilde{\mathbf{Z}}\mathbf{c} + \text{résidus} \quad (6.23)$$

et de calculer un test en  $F$  pour  $\mathbf{c} = \mathbf{0}$  ou une des statistiques de test équivalente. De manière implicite bien sûr,  $\tilde{\mathbf{Z}}$  doit correspondre à la matrice  $\tilde{\mathbf{X}}_2$  pour un modèle non contraint *quelconque* qui comprend (6.22) comme un cas particulier. Mais il y a des situations pour lesquelles il est naturel de dériver la GNR (6.23) sans préciser explicitement un tel modèle, pour évaluer un **test diagnostique**. De tels tests sont abondamment utilisés lorsque l'on doit

estimer un modèle et que l'on désire savoir s'il y a une quelconque évidence sur le fait qu'il soit mal spécifié. Nous allons rencontrer de tels tests dans la section qui suit et tout au long du livre.

## 6.5 TESTS DIAGNOSTIQUES POUR LES RÉGRESSIONS LINÉAIRES

Les résultats précédents sur l'usage de la GNR pour tester des contraintes portant sur les paramètres des modèles de régression non linéaire sont bien sûr transposables aux modèles de régression linéaire. Il est utile de considérer en peu de mots le cas de ces derniers, d'une part parce qu'une grande part de la littérature y est consacrée, d'autre part parce que cela nous donnera l'opportunité de discuter des tests diagnostiques (qui sont souvent, à tort, considérés comme différents des autres tests portant sur des contraintes), et enfin parce qu'il y a toujours du mérite à envisager les cas les plus simples possibles.

Supposons que l'on ait estimé le modèle de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (6.24)$$

où  $\mathbf{X}$  est une matrice de dimension  $n \times k$  et  $\boldsymbol{\beta}$  est un vecteur à  $k$  composantes, et que l'on veuille le tester sur de mauvaises descriptions éventuelles de la fonction de régression. Nous supposons la normalité pour l'instant afin de pouvoir discuter sur la base de tests exacts. Souvenons-nous que la fonction de régression est supposée déterminer  $E(\mathbf{y} | \Omega)$ , l'espérance de  $\mathbf{y}$  conditionnée par un ensemble d'informations donné  $\Omega$ . Soit  $\mathbf{Z}$  une matrice de dimension  $n \times l$  contenant les observations de n'importe quel ensemble de régresseurs qui appartiennent à  $\Omega$  mais qui ne se trouvent pas dans  $\mathcal{S}(\mathbf{X})$ . Alors, si l'hypothèse nulle  $E(\mathbf{y} | \Omega) = \mathbf{X}\boldsymbol{\beta}$  est exacte, l'estimation du vecteur  $\boldsymbol{\gamma}$  dans la régression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u} \quad (6.25)$$

devrait être très peu différente de zéro. Bien sûr, il serait possible de tester cette hypothèse en calculant un Fisher ordinaire pour  $\boldsymbol{\gamma} = \mathbf{0}$ :

$$\frac{(\text{RSSR} - \text{USSR})/l}{\text{USSR}/(n - k - l)}, \quad (6.26)$$

où RSSR et USSR sont respectivement les sommes des résidus au carré de (6.24) et (6.25). Si la valeur de la statistique de test est forte (et donc le  $P$  associé est faible), nous voudrions rejeter l'hypothèse nulle, et conclure ainsi que le modèle (6.24) est mal spécifié. Ceci est un exemple de ce que certains auteurs (dont Pagan (1984a) et Pagan et Hall (1983)) appellent les **tests à variable additionnelle**. En français, on parle aussi d'une **régression augmentée**.



Que se passerait-il si l'on utilisait une régression de Gauss-Newton au lieu de la régression plus naturelle (6.25)? La GNR (6.17) devient

$$\mathbf{M}_X \mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{c} + \text{résidus}, \quad (6.27)$$

qui grâce au Théorème FWL entraîne la même SSR que la régression

$$\mathbf{M}_X \mathbf{y} = \mathbf{M}_X \mathbf{Z}\mathbf{c} + \text{résidus}. \quad (6.28)$$

Mais si l'on applique le Théorème FWL à la régression (6.25), nous constatons que la SSR de cette régression est identique à la SSR de la régression (6.28). Ainsi, dans cette situation le Fisher basé sur la régression de Gauss-Newton (6.27) sera identique à (6.29), le Fisher basé sur la régression ordinaire (6.25). Nous voyons que les tests fondés sur la GNR sont équivalents aux tests à variable additionnelle lorsque ceux-ci sont applicables.

Il semble souvent intéressant de fonder des tests portant sur une mauvaise spécification d'un modèle sur les résidus  $\hat{\mathbf{u}}$ , car ils fournissent des estimations des aléas  $\mathbf{u}$  (Pagan et Hall (1983) montrent la façon de construire un grand nombre de tests qui vont dans ce sens). Ainsi, il semble naturel de tester la qualité de la spécification en régressant simplement les résidus  $\hat{\mathbf{u}} = \mathbf{M}_X \mathbf{y}$  sur les régresseurs de tests  $\mathbf{Z}$ .

L'hypothèse que  $E(\mathbf{y}|\Omega) = \mathbf{X}\boldsymbol{\beta}$  peut être testée contre toute autre spécification alternative de l'espérance conditionnelle en testant l'importance d'une quelconque matrice de test  $\mathbf{Z}$  dans les régressions (6.25) et (6.27). La seule chose nécessaire est que  $\mathbf{Z}$  soit asymptotiquement non corrélée à  $\mathbf{u}$  et ne dépende de rien d'autre que ce qui fait partie de l'ensemble d'informations  $\Omega$ . Un grand nombre de tests de spécification de ce genre a été proposé et nous en rencontrerons beaucoup au cours de cet ouvrage. Un exemple bien connu est le **test d'erreur de spécification de la régression** (en anglais **RESET**, l'abréviation que nous emploierons) qui a été présenté la première fois par Ramsey (1969) (voir également Anscombe (1961)), et remanié sous la forme d'un test en  $F$  pour les variables omises, par Ramsey et Schmidt (1976). Pour ce test chaque colonne de  $\mathbf{Z}$  est composée d'une puissance des valeurs ajustées  $\mathbf{X}\hat{\boldsymbol{\beta}}$ , telle que le carré des valeurs ajustées, le cube des valeurs ajustées, et ainsi de suite. Dans le cas le plus simple, il n'y a qu'un régresseur de test, qui est le régresseur des valeurs ajustées au carré, et cette version épurée du test RESET est souvent la plus utile. Il est intéressant de constater qu'il peut se dériver comme une application de la régression de Gauss-Newton.

Supposons que le modèle soumis au test soit une fois encore (6.24) que l'on désire tester contre l'hypothèse alternative explicite

$$y_t = \mathbf{X}_t \boldsymbol{\beta} (1 + \theta \mathbf{X}_t \boldsymbol{\beta}) + u_t, \quad (6.29)$$

où  $\theta$  est un paramètre inconnu qu'il faut évaluer. Lorsque  $\theta = 0$ , ce modèle se résume à (6.24), mais lorsque  $\theta$  est non nul il tient compte d'une relation

non linéaire entre  $\mathbf{X}_t$  et  $y_t$ . Beaucoup d'autres modèles non linéaires seraient approximés avec satisfaction par (6.29) au voisinage de  $\theta = 0$ , de sorte que cela semble être une alternative raisonnable. Il est aisé de voir que la GNR correspondante de (6.29) est

$$y_t - \mathbf{X}_t\boldsymbol{\beta}(1 + \theta\mathbf{X}_t\boldsymbol{\beta}) = (2\theta(\mathbf{X}_t\boldsymbol{\beta})\mathbf{X}_t + \mathbf{X}_t)\mathbf{b} + (\mathbf{X}_t\boldsymbol{\beta})^2c + \text{résidu}.$$

Lorsque cette expression est évaluée en  $\hat{\boldsymbol{\beta}}$ , l'estimation OLS sous l'hypothèse nulle  $\theta = 0$ , elle se résume à

$$y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}} = \mathbf{X}_t\mathbf{b} + (\mathbf{X}_t\hat{\boldsymbol{\beta}})^2c + \text{résidu}, \quad (6.30)$$

et ainsi que nous l'avons vu le Student de  $\hat{c}$  dans cette GNR sera identique au Student de l'estimation de  $c$  dans la régression

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + (\mathbf{X}_t\hat{\boldsymbol{\beta}})^2c + \text{résidu},$$

qui est la régression nécessaire à l'exécution de la version la plus simple du test RESET. Ainsi le test RESET fournit un moyen simple de tester la non linéarité dans la relation entre  $\mathbf{X}$  et  $\mathbf{y}$ ; pour davantage de détails à ce sujet, consulter MacKinnon et Magee (1990). Ce test est évidemment transposable aux modèles de régression non linéaire. Si le modèle soumis au test était  $y_t = x_t(\boldsymbol{\beta}) + u_t$ , il suffirait de remplacer  $\mathbf{X}_t\hat{\boldsymbol{\beta}}$  par  $\hat{x}_t$  deux fois, lorsqu'elle apparaît dans la régression (6.30), afin d'obtenir une GNR adéquate.

## 6.6 ESTIMATION EFFICACE EN UNE ÉTAPE

Il est quelquefois facile d'obtenir des estimations convergentes mais inefficaces, mais il est relativement délicat d'obtenir des estimations NLS. Cela peut être le cas, lorsque par exemple, le modèle non linéaire qu'il faut estimer est en réalité un modèle linéaire soumis à des contraintes non linéaires, comme c'est le cas avec de nombreux modèles d'anticipations rationnelles. Dans de telles circonstances, un résultat utile est que si l'on franchit *une seule* étape à partir de ces estimations convergentes initiales, en utilisant la régression de Gauss-Newton, les estimations que l'on obtiendra seront équivalentes asymptotiquement aux estimations NLS.

Si  $\hat{\boldsymbol{\beta}}$  désigne les estimations initiales qui sont supposées être convergentes à un taux  $n^{1/2}$ . La GNR est alors

$$\mathbf{y} - \hat{\mathbf{x}} = \hat{\mathbf{X}}\mathbf{b} + \text{résidus},$$

et l'estimation de  $\mathbf{b}$  de cette régression est

$$\hat{\mathbf{b}} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top (\mathbf{y} - \hat{\mathbf{x}}). \quad (6.31)$$

Alors l'estimateur en une étape est

$$\dot{\beta} = \hat{\beta} + \hat{b}.$$

Un développement de Taylor de  $\mathbf{x}(\dot{\beta})$  autour de  $\beta = \beta_0$  entraîne

$$\dot{\mathbf{x}} \cong \mathbf{x}_0 + \mathbf{X}_0(\dot{\beta} - \beta_0),$$

où  $\mathbf{x}_0 \equiv \mathbf{x}(\beta_0)$  et  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$ . En mettant ce résultat dans (6.31), en remplaçant  $\mathbf{y}$  par sa valeur sous le DGP,  $\mathbf{x}_0 + \mathbf{u}$ , et en introduisant des puissances adéquates de  $n$  de sorte que toutes les quantités soient  $O(1)$ , il vient le résultat suivant

$$\begin{aligned} n^{1/2}\dot{\mathbf{b}} &\cong n^{-1/2}(n^{-1}\dot{\mathbf{X}}^\top\dot{\mathbf{X}})^{-1}\dot{\mathbf{X}}^\top(\mathbf{x}_0 + \mathbf{u} - \mathbf{x}_0 - \mathbf{X}_0(\dot{\beta} - \beta_0)) \\ &= (n^{-1}\dot{\mathbf{X}}^\top\dot{\mathbf{X}})^{-1}(n^{-1/2}\dot{\mathbf{X}}^\top\mathbf{u} - (n^{-1}\dot{\mathbf{X}}^\top\mathbf{X}_0)n^{1/2}(\dot{\beta} - \beta_0)). \end{aligned}$$

Mais notons que

$$n^{-1}\dot{\mathbf{X}}^\top\dot{\mathbf{X}} \stackrel{a}{=} n^{-1}\mathbf{X}_0^\top\mathbf{X}_0 \stackrel{a}{=} n^{-1}\dot{\mathbf{X}}^\top\mathbf{X}_0,$$

qui découle de la convergence de  $\dot{\beta}$ . Ainsi,

$$n^{1/2}\dot{\mathbf{b}} \stackrel{a}{=} (n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}(n^{-1/2}\mathbf{X}_0^\top\mathbf{u}) - n^{1/2}(\dot{\beta} - \beta_0).$$

En ajoutant cette expression à  $n^{1/2}\dot{\beta}$  afin d'obtenir  $n^{1/2}$  fois l'estimateur en une étape  $\dot{\beta}$ , nous constatons que

$$n^{1/2}(\dot{\beta} - \beta_0) \cong (n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}(n^{-1/2}\mathbf{X}_0^\top\mathbf{u}).$$

En prenant la limite en probabilité de  $n^{-1}\mathbf{X}_0^\top\mathbf{X}_0$ , cela devient

$$n^{1/2}(\dot{\beta} - \beta_0) \stackrel{a}{=} \text{plim}_{n \rightarrow \infty} (n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}(n^{-1/2}\mathbf{X}_0^\top\mathbf{u}). \quad (6.32)$$

Le membre de droite de cette expression nous semble familier. En réalité, le résultat (5.39) qui provient du Chapitre 5 nous montre que  $n^{1/2}(\hat{\beta} - \beta_0)$  est asymptotiquement égal au membre de droite de (6.32). Nous avons ainsi démontré que l'estimateur en une étape  $\dot{\beta}$  est asymptotiquement équivalent à l'estimateur NLS  $\hat{\beta}$ . Par conséquent, il doit avoir la même distribution asymptotique que  $\hat{\beta}$ , et nous pouvons donc conclure que

$$n^{1/2}(\dot{\beta} - \beta_0) \stackrel{a}{\sim} N(\mathbf{0}, \sigma^2(n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}).$$

Une estimation en une étape peut se révéler particulièrement utile pour imposer des contraintes non linéaires à un modèle qu'il est aisé d'estimer sans contraintes mais qui devient plus délicat à évaluer lorsqu'il est soumis à

des restrictions. En particulier, supposons que la fonction de régression non contrainte soit  $\mathbf{X}\boldsymbol{\beta}$  et que la fonction de régression contrainte puisse s'écrire sous la forme  $\mathbf{X}\boldsymbol{\beta}(\boldsymbol{\gamma})$ , où  $\boldsymbol{\beta}(\boldsymbol{\gamma})$  est un vecteur dont les  $k$  composantes sont des fonctions du vecteur  $\boldsymbol{\gamma}$  à  $l$  composantes,  $l$  étant inférieur à  $k$ , de façon à ce que le modèle contraint soit non linéaire en ses paramètres uniquement. Quelques éléments de  $\boldsymbol{\beta}(\boldsymbol{\gamma})$  peuvent évidemment être nuls. Dans ce cas le modèle contraint est

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}(\boldsymbol{\gamma}) + \mathbf{u}, \quad (6.33)$$

et le modèle non contraint est

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \quad (6.34)$$

Le vecteur des estimations OLS de (6.34),  $\hat{\boldsymbol{\beta}}$ , fournit un vecteur d'estimations initiales convergentes  $\hat{\boldsymbol{\gamma}}$ . Celui-ci peut être facile ou non à calculer, et sera certainement non unique, puisqu'il y a moins d'éléments dans  $\boldsymbol{\gamma}$  que dans  $\boldsymbol{\beta}$ . La régression qui doit servir à l'obtention des estimations en une étape est la GNR correspondant à (6.33) avec le vecteur de paramètres  $\boldsymbol{\gamma}$  évalué en  $\hat{\boldsymbol{\gamma}}$ :

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta}(\hat{\boldsymbol{\gamma}}) = \hat{\mathbf{X}}^* \mathbf{c} + \text{résidus},$$

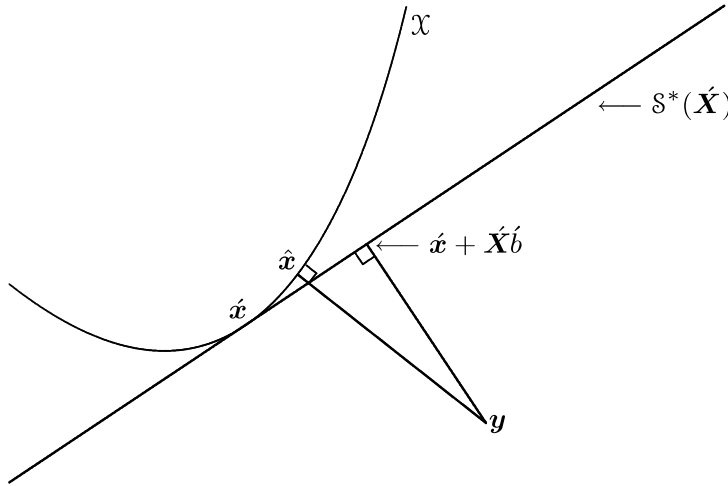
où la matrice  $\hat{\mathbf{X}}^*$  de dimension  $n \times l$  est définie par

$$\hat{\mathbf{X}}^* \equiv \mathbf{X} \left. \frac{\partial \boldsymbol{\beta}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \right|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}}.$$

Comme d'habitude les estimations en une étape sont  $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}} + \hat{\mathbf{c}}$ , et elles seront asymptotiquement équivalentes aux estimations contraintes  $\tilde{\boldsymbol{\gamma}}$ , qui seraient de loin plus coûteuses à obtenir.

Intuitivement, les estimateurs efficaces en une étape basés sur la GNR sont asymptotiquement équivalents aux estimateurs NLS pour une raison très semblable à celle que nous avons exposée à la section précédente pour la validité des tests fondés sur la régression de Gauss-Newton. La GNR réalise la minimisation de  $SSR^*(\hat{\boldsymbol{\beta}} + \mathbf{b})$ , qui est une approximation quadratique de  $SSR(\boldsymbol{\beta})$  autour de  $\hat{\boldsymbol{\beta}}$ . Asymptotiquement, la fonction  $SSR(\boldsymbol{\beta})$  est quadratique au voisinage de  $\boldsymbol{\beta}_0$ . Lorsque la taille d'échantillon est suffisamment importante, la convergence de  $\hat{\boldsymbol{\beta}}$  implique que nous prenions l'approximation quadratique en un point proche de  $\boldsymbol{\beta}_0$ , et ainsi l'approximation coïncidera asymptotiquement avec  $SSR(\boldsymbol{\beta})$  elle-même.

Nous pouvons constater ce qu'il advient en considérant la Figure 6.4. Comme dans les figures de ce genre que nous avons déjà vues (par exemple la Figure 2.2),  $k = 1$  et nous supposons que  $\mathbf{x}(\boldsymbol{\beta})$  se situe, du moins localement, dans un sous-espace bi-dimensionnel de  $\mathbb{R}^n$ , ce qui nous permet de la tracer sur le papier. C'est délibérément que nous l'avons dessinée avec un fort degré de non linéarité dans cet espace, en vue de rendre les résultats plus faciles à



**Figure 6.4** Estimation efficace en une étape

assimiler. La variété  $\mathcal{X}$  dessine ainsi une ligne fortement incurvée dans  $\mathbb{R}^n$ . Au point  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\boldsymbol{\beta}})$  nous prenons une approximation linéaire,  $\hat{\mathbf{x}} + \hat{\mathbf{X}}\hat{\mathbf{b}}$ . Le sous-espace engendré par les colonnes de  $\hat{\mathbf{X}}$ , translaté de manière à ce qu'il soit tangent à la variété  $\mathcal{X}$  en  $\hat{\mathbf{x}}$ , est désigné par  $\mathcal{S}^*(\hat{\mathbf{X}})$ . L'estimation en une étape implique la projection orthogonale de  $\mathbf{y}$  sur ce sous-espace, afin d'obtenir une estimation du coefficient,  $\hat{\mathbf{b}}$ . Nous obtenons à présent l'estimation en une étape  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \hat{\mathbf{b}}$ . Au contraire, les moindres carrés non linéaires impliquent la projection orthogonale de  $\mathbf{y}$  sur  $\mathcal{X}$ , au point  $\hat{\mathbf{x}}$ . À l'évidence  $\hat{\boldsymbol{\beta}}$  et  $\hat{\boldsymbol{\beta}}$  sont généralement différents à moins que  $\mathcal{S}^*(\hat{\mathbf{X}})$  ne coïncide avec  $\mathcal{X}$  au voisinage de  $\hat{\boldsymbol{\beta}}$ . Mais les estimations NLS et les estimations en une étape n'en sont pas moins asymptotiquement équivalentes parce que la convergence de  $\hat{\mathbf{x}}$  implique que, asymptotiquement, il est si proche de  $\hat{\mathbf{x}}$  que  $\mathcal{X}$  ne peut pas être suffisamment courbée entre les deux points.

Malheureusement le fait que les estimateurs en une étape fondés sur la GNR soient asymptotiquement équivalents aux estimateurs NLS n'implique pas que ceux-ci possèdent les mêmes propriétés que ces derniers en échantillons finis. Souvenons-nous que l'équivalence implique que le DGP soit  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0) + \mathbf{u}$ . Si le DGP n'est pas un cas particulier du modèle, l'équivalence est caduque. Même dans le cas contraire, un grand nombre de choses dépend de la qualité de l'estimateur convergent initial  $\hat{\boldsymbol{\beta}}$ . Lorsque  $\hat{\boldsymbol{\beta}}$  est proche de  $\boldsymbol{\beta}_0$  et que l'échantillon a une taille importante,  $SSR^*(\hat{\boldsymbol{\beta}} + \mathbf{b})$  devrait se révéler être une très bonne approximation de  $SSR(\boldsymbol{\beta})$ , et donc  $\hat{\boldsymbol{\beta}}$  devrait être très près de l'estimation NLS  $\hat{\boldsymbol{\beta}}$ . D'un autre côté lorsque l'estimation convergente est éloignée de  $\boldsymbol{\beta}_0$  (et le fait qu'un estimateur soit convergent n'empêche pas qu'il soit extrêmement peu efficace), les estimations en une étape peuvent différer fortement des estimations NLS. Lorsque la différence entre les deux est significative, nous recommanderions l'usage des estimations NLS, bien qu'en l'absence d'une étude détaillée du modèle en cause, il ne soit pas possi-

ble d'être catégorique sur le choix de l'usage d'un estimateur plutôt que d'un autre lorsque les deux sont asymptotiquement équivalents. Les estimations en une étape prennent davantage de sens lorsque la taille de l'échantillon est importante, ce qui implique que l'estimateur convergent initial est vraisemblablement satisfaisant, et aussi que les moindres carrés non linéaires puissent être coûteux à employer.

## 6.7 TESTS BASÉS SUR UNE ESTIMATION CONVERGENTE

Les procédures de test dont nous avons discuté dans les Sections 6.4 et 6.5 impliquent toutes l'évaluation d'une régression artificielle aux estimations NLS contraintes, et entraînent par conséquent des statistiques de test basées sur le principe LM. Mais lorsque la fonction de régression contrainte est non linéaire, il n'est pas toujours pratique d'obtenir des estimations NLS. Par chance, il est toujours possible d'exécuter des tests au moyen d'une GNR lorsque n'importe quelle estimation convergente à un taux  $n^{1/2}$  qui satisfait l'hypothèse nulle est disponible. Dans cette section nous discutons brièvement de la façon de procéder.

Supposons que l'on traite la situation dont nous avons discuté à la Section 6.4, dans laquelle le vecteur de paramètres  $\beta$  est partitionné en  $[\beta_1 : \beta_2]$ , et l'hypothèse nulle est  $\beta_2 = \mathbf{0}$ . Supposons que l'on dispose d'un vecteur d'estimations convergentes à un taux  $n^{1/2}$   $\hat{\beta} \equiv [\hat{\beta}_1 : \mathbf{0}]$ . Alors la GNR, exprimée dans la notation appropriée, est

$$\mathbf{y} - \hat{\mathbf{x}} = \hat{\mathbf{X}}_1 \mathbf{b}_1 + \hat{\mathbf{X}}_2 \mathbf{b}_2 + \text{résidus}. \quad (6.35)$$

La somme des carrés expliqués de cette régression est

$$(\mathbf{y} - \hat{\mathbf{x}})^\top \hat{\mathbf{P}}_1 (\mathbf{y} - \hat{\mathbf{x}}) + (\mathbf{y} - \hat{\mathbf{x}})^\top \hat{\mathbf{M}}_1 \hat{\mathbf{X}}_2 (\hat{\mathbf{X}}_2^\top \hat{\mathbf{M}}_1 \hat{\mathbf{X}}_2)^{-1} \hat{\mathbf{X}}_2^\top \hat{\mathbf{M}}_1 (\mathbf{y} - \hat{\mathbf{x}}). \quad (6.36)$$

Le premier terme est ici la somme des carrés expliqués de la régression de  $\mathbf{y} - \hat{\mathbf{x}}$  sur  $\hat{\mathbf{X}}_1$  uniquement, et le second terme représente l'accroissement de la somme des carrés expliqués causé par la prise en compte de  $\hat{\mathbf{X}}_2$ . Remarquons que le premier terme est en général non nul, parce que  $\hat{\beta}_1$  ne satisfera pas en général les conditions du premier ordre pour les estimations NLS du modèle contraint.

La différence entre la somme des carrés expliqués de (6.35) et la somme des carrés expliqués de la régression de  $\mathbf{y} - \hat{\mathbf{x}}$  sur  $\hat{\mathbf{X}}_1$  uniquement est

$$(\mathbf{y} - \hat{\mathbf{x}})^\top \hat{\mathbf{M}}_1 \hat{\mathbf{X}}_2 (\hat{\mathbf{X}}_2^\top \hat{\mathbf{M}}_1 \hat{\mathbf{X}}_2)^{-1} \hat{\mathbf{X}}_2^\top \hat{\mathbf{M}}_1 (\mathbf{y} - \hat{\mathbf{x}}) = \|\mathbf{P}_{\hat{\mathbf{M}}_1 \hat{\mathbf{X}}_2} (\mathbf{y} - \hat{\mathbf{x}})\|^2.$$

Cela ressemble justement au numérateur du Fisher (6.24). En réalité, l'unique différence est que tout est évalué avec les estimations convergentes au taux  $n^{1/2}$   $\hat{\beta}$ , au lieu des estimations NLS contraintes  $\tilde{\beta}$ . Il ne devrait donc pas

être surprenant d'apprendre que le Fisher pour  $\mathbf{b}_2 = \mathbf{0}$  est asymptotiquement équivalent (sous l'hypothèse nulle) à celui pour  $\mathbf{b}_2 = \mathbf{0}$  dans le test LM plus conventionnel de la GNR (6.17).

Nous ne nous emploierons pas à démontrer ce résultat formellement. L'intuition est tellement évidente qu'une démonstration est à peine nécessaire. D'après les résultats des sections qui ont précédé, nous savons que

$$n^{1/2}(\hat{\boldsymbol{\beta}} + \hat{\mathbf{b}} - \boldsymbol{\beta}_0) \stackrel{a}{=} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

où  $\hat{\mathbf{b}}$  est l'estimation OLS de  $\mathbf{b}$  dans (6.35) et  $\hat{\boldsymbol{\beta}}$  est l'estimation NLS non contrainte. Ainsi  $\hat{\boldsymbol{\beta}} + \hat{\mathbf{b}}$  est un vecteur composé des estimations convergentes en une étape. Puisque  $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ , nous avons

$$n^{1/2}(\hat{\mathbf{b}}_2 - \boldsymbol{\beta}_2^0) \stackrel{a}{=} n^{1/2}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^0),$$

où  $\boldsymbol{\beta}_2^0$  est la valeur de  $\boldsymbol{\beta}_2$  sous le DGP. Ainsi, l'estimation OLS de  $\mathbf{b}_2$  dans (6.35) est asymptotiquement équivalente à l'estimation NLS non contrainte  $\hat{\boldsymbol{\beta}}_2$ . Ceci devrait rendre intuitivement évident qu'un test en  $F$  pour  $\mathbf{b}_2 = \mathbf{0}$  dans (6.35) est équivalent à un test pour  $\boldsymbol{\beta}_2 = \mathbf{0}$ .

Nous pouvons calculer des tests pour  $\boldsymbol{\beta}_2 = \mathbf{0}$  en utilisant la régression (6.35) de la même façon qu'en utilisant (6.17), avec toutefois une exception. La quantité  $nR^2$  dans (6.17) est une statistique de test valide, mais la même quantité dans (6.35) n'en est pas une. La raison en est que  $\hat{\mathbf{X}}_1$  aura généralement quelque capacité explicative sur la variation de  $\mathbf{y} - \hat{\mathbf{x}}$ , de sorte que le premier terme dans (6.36) sera non nul. Nous pourrions construire une statistique de test valide comme  $nR^2$  dans (6.35) moins  $nR^2$  dans la régression de  $\mathbf{y} - \hat{\mathbf{x}}$  sur  $\hat{\mathbf{X}}_1$  uniquement. Cependant, il est préférable d'utiliser simplement des tests en  $F$  ou en  $t$ , qui sont calculés comme si (6.35) était une régression naturelle plutôt qu'une régression artificielle.

Dans la littérature consacrée à l'estimation par maximum de vraisemblance, les tests basés sur le choix arbitraire des estimateurs convergents à un taux  $n^{1/2}$  sont appelés des tests  $\mathbf{C}(\boldsymbol{\alpha})$ . De tels tests ont été proposés à l'initiative de Neyman (1959); pour plus de références et de détails, consulter la Section 13.7. Comme nous le verrons à ce moment, il est possible d'interpréter les tests dont nous avons discuté dans cette section comme des tests  $\mathbf{C}(\boldsymbol{\alpha})$ . Mais ces tests pourraient être aussi interprétés comme des tests de Wald dans certains cas. Supposons que  $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_1 \vdots \mathbf{0}]$ , où  $\hat{\boldsymbol{\beta}}_1$  est l'estimation NLS non contrainte de  $\boldsymbol{\beta}_1$ . Ce choix pour  $\hat{\boldsymbol{\beta}}$  est certainement convergent à un taux  $n^{1/2}$  et satisfait l'hypothèse nulle, et par conséquent entraînera clairement des statistiques de test valides. Puisque la GNR dépend uniquement des estimations non contraintes  $\hat{\boldsymbol{\beta}}$ , nous nous référerons aux tests calculés de cette manière en tant que **pseudo-Wald**, bien qu'ils ne soient pas fondés sur le principe de Wald. De tels tests pseudo-Wald peuvent s'avérer plus faciles à calculer que les tests de Wald plus conventionnels.

## 6.8 ESTIMATION NON LINÉAIRE UTILISANT LA GNR

Nous discutons dans cette section de la manière d'utiliser la régression de Gauss-Newton comme une partie d'un algorithme efficace de minimisation des fonctions somme des carrés. C'était en réalité le but originel de la GNR. Le terme "Gauss-Newton" est en fait emprunté à la littérature consacrée à l'optimisation numérique appliquée aux problèmes des moindres carrés non linéaires, et la plupart des autres usages de cette régression artificielle sont relativement récents comme nous allons le voir dans la Section 6.9 qui suit.

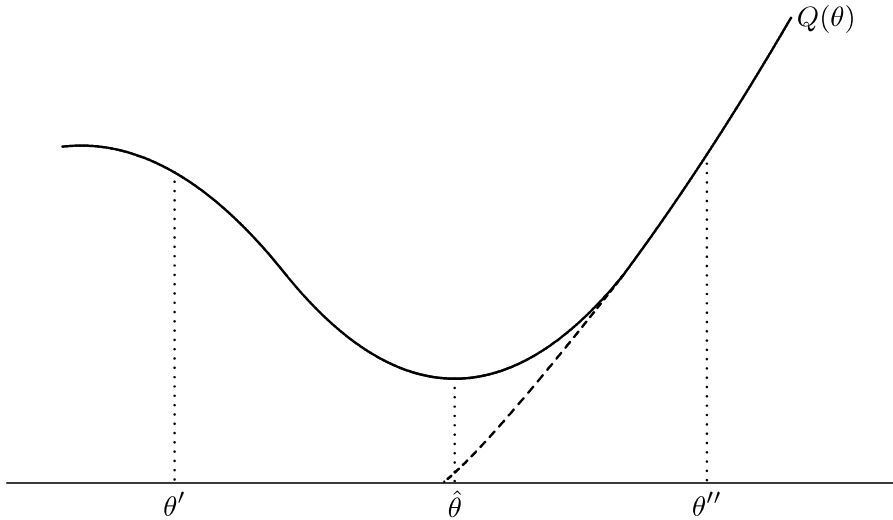
La grande majorité des algorithmes efficaces qui ont pour but la maximisation ou la minimisation d'une fonction à deux ou davantage de variables, disons  $Q(\boldsymbol{\theta})$ , procèdent sensiblement de la même manière. Un tel algorithme entame une série d'itérations; pour chacune desquelles il débute avec une valeur particulière de  $\boldsymbol{\theta}$ , disons  $\boldsymbol{\theta}^{(j)}$ , et tente de trouver une valeur plus satisfaisante. L'algorithme choisit tout d'abord une direction de recherche de la meilleure valeur de  $\boldsymbol{\theta}$ , et décide ensuite de la distance à parcourir dans cette direction. La différence majeure entre les algorithmes d'optimisation non contrainte réside dans la manière de choisir la direction de recherche, et dans la manière de déterminer la taille du dernier mouvement dans cette direction. L'éventail des choix est large.

Remarquons que n'importe quel algorithme de minimisation peut tout aussi bien être utilisé pour la maximisation, puisque minimiser  $Q(\boldsymbol{\theta})$  est équivalent à maximiser  $-Q(\boldsymbol{\theta})$ . En adoptant les conventions en usage dans la littérature, nous traiterons le cas de la minimisation, ce qui de toute façon est ce qui nous intéresse avec la fonction somme des carrés.<sup>2</sup> Dans cette section, nous tenterons de donner une vue d'ensemble de la "mécanique" des algorithmes de minimisation numériques, et de la façon dont on peut employer la régression de Gauss-Newton dans ces algorithmes, mais nous ne discuterons pas des conclusions importantes relatives à l'ordinateur qui affectent substantiellement les performances des algorithmes informatiques. Une référence remarquable concernant l'art et la science de l'optimisation numérique est Gill, Murray, et Wright (1981); consulter également Bard (1974), Quandt (1983), Press, Flannery, Teukolsky, et Vetterling (1986, Chapitre 10), et Seber et Wild (1989, Chapitre 14).

Une des techniques fondamentales de l'optimisation numérique est la **Méthode de Newton**. Supposons que l'on veuille minimiser une fonction non linéaire  $Q(\boldsymbol{\theta})$ , où  $\boldsymbol{\theta}$  est un vecteur à  $k$  éléments. Etant donnée n'importe quelle valeur initiale, disons  $\boldsymbol{\theta}^{(1)}$ , nous obtenons une approximation de  $Q(\boldsymbol{\theta})$  par un

<sup>2</sup> Cependant, lorsque nous traitons des fonctions de vraisemblance, nous voudrions les maximiser (voir le Chapitre 8). On peut donc appliquer la discussion qui va suivre à de tels cas, moyennant quelques modifications mineures.





**Figure 6.5** Cas pour lesquels la méthode de Newton ne sera pas performante

développement en série de Taylor à l'ordre un autour de  $\theta^{(1)}$ :

$$\begin{aligned} Q^*(\theta) &= Q(\theta^{(1)}) + (\mathbf{g}^{(1)})^\top (\theta - \theta^{(1)}) + \frac{1}{2} (\theta - \theta^{(1)})^\top \mathbf{H}^{(1)} (\theta - \theta^{(1)}) \\ &\cong Q(\theta), \end{aligned}$$

où  $\mathbf{g}(\theta)$ , le gradient de  $Q(\theta)$ , est un vecteur colonne à  $k$  éléments dont la composante type est  $\partial Q(\theta)/\partial \theta_i$ , et  $\mathbf{H}(\theta)$ , la matrice Hessienne de  $Q(\theta)$ , est une matrice de dimension  $k \times k$  dont l'élément type est  $\partial^2 Q(\theta)/\partial \theta_i \partial \theta_l$ ;  $\mathbf{g}^{(1)}$  et  $\mathbf{H}^{(1)}$  correspondant à  $\mathbf{g}(\theta^{(1)})$  et  $\mathbf{H}(\theta^{(1)})$ . La résolution des conditions du premier ordre pour un minimum de  $Q^*(\theta)$  par rapport à  $\theta$  entraîne une nouvelle valeur de  $\theta$ , que nous appellerons  $\theta^{(2)}$ . Elle dépend dans une relation simple de  $\theta^{(1)}$ , et du gradient et de la matrice Hessienne évalués en  $\theta^{(1)}$ :

$$\theta^{(2)} = \theta^{(1)} - (\mathbf{H}^{(1)})^{-1} \mathbf{g}^{(1)}. \quad (6.37)$$

L'équation (6.37) est la clef de la Méthode de Newton. Si l'approximation quadratique  $Q^*(\theta)$  est une fonction strictement convexe, ce qui sera le cas si et seulement si la matrice Hessienne  $\mathbf{H}(\theta^{(1)})$  est définie positive,  $\theta^{(2)}$  sera le minimum global de  $Q^*(\theta)$ . Si de plus  $Q^*(\theta)$  est une bonne approximation de  $Q(\theta)$ ,  $\theta^{(2)}$  devrait être proche de  $\hat{\theta}$ , le minimum de  $Q(\theta)$ . La Méthode de Newton implique l'usage répété de l'équation (6.37) pour avoir une série de valeurs  $\theta^{(2)}, \theta^{(3)}, \dots$ . Lorsque la fonction originelle  $Q(\theta)$  est quadratique et possède un minimum global en  $\hat{\theta}$ , la Méthode de Newton décèle clairement  $\hat{\theta}$  dès la première étape, puisque l'approximation quadratique est exacte. Lorsque  $Q(\theta)$  est approximativement quadratique, comme le sont toutes les fonctions somme des carrés lorsque l'on est suffisamment proche de leurs minima, la Méthode de Newton converge généralement très rapidement.

Dans de nombreux autres cas toutefois, la Méthode de Newton échoue totalement, en particulier si  $Q(\boldsymbol{\theta})$  n'est pas convexe au voisinage de  $\boldsymbol{\theta}^{(j)}$  pour tout  $j$  appartenant à la série des  $\boldsymbol{\theta}^{(i)}$ . Pour s'en persuader, considérons la Figure 6.5. La fonction unidimensionnelle que nous avons représentée possède un minimum global en  $\hat{\theta}$ , mais lorsque la Méthode de Newton démarre en des points tels que  $\theta'$  ou  $\theta''$ , elle ne peut pas trouver  $\hat{\theta}$ . Dans le premier cas de figure,  $Q(\theta)$  est concave en  $\theta'$  au lieu d'être convexe, de sorte que la Méthode de Newton détourne le calcul dans la mauvaise direction. Dans le second cas de figure, l'approximation quadratique en  $\theta''$  est très peu satisfaisante pour des valeurs éloignées de  $\theta''$ , parce que  $Q(\theta)$  est aplatie aux environs de  $\theta''$ . L'approximation quadratique  $Q^*(\theta)$  de  $Q(\theta)$  considérée en  $\theta''$  est représentée pour la courbe en tirets. Elle aura bien sûr un minimum situé à gauche de  $\hat{\theta}$ . Néanmoins, la plupart des techniques efficaces d'optimisation non linéaire pour des problèmes "faciles" sont des versions modifiées de la Méthode de Newton, qui visent à retenir ses points forts tout en essayant de surmonter des difficultés telles que celles illustrées à la Figure 6.5.

Les techniques de minimisation numérique qui sont basées sur la Méthode de Newton remplacent (6.37) par une formule légèrement plus compliquée

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \alpha^{(j)} (\mathbf{D}^{(j)})^{-1} \mathbf{g}^{(j)}, \quad (6.38)$$

qui définit  $\boldsymbol{\theta}^{(j+1)}$ , la valeur de  $\boldsymbol{\theta}$  à l'étape  $j + 1$ , comme une fonction de  $\boldsymbol{\theta}^{(j)}$ . Ici  $\alpha^{(j)}$  est un scalaire qui est habituellement déterminé de manière endogène, au fur et à mesure du calcul de l'algorithme, et  $\mathbf{D}^{(j)} \equiv \mathbf{D}(\boldsymbol{\theta}^{(j)})$  est une matrice qui approxime  $\mathbf{H}(\boldsymbol{\theta}^{(j)})$  aux environs du minimum mais qui est toujours construite de telle sorte qu'elle est toujours définie positive. La plupart de ces algorithmes consistent en une série de deux étapes alternées. En partant de  $\boldsymbol{\theta}^{(j)}$  ils calculent tout d'abord  $\mathbf{g}^{(j)}$  et  $\mathbf{D}^{(j)}$  et déterminent ensuite quelle direction choisir. Puis ils résolvent un problème de minimisation à une dimension pour trouver  $\alpha^{(j)}$ , qui doit indiquer la distance à parcourir dans cette direction. Ces deux étapes doivent simultanément donner  $\boldsymbol{\theta}^{(j+1)}$ . Les algorithmes reviennent ensuite une étape en arrière et continuent à alterner les deux étapes jusqu'à ce qu'ils décident que l'approximation de  $\hat{\boldsymbol{\theta}}$  est suffisamment fine.

Parce qu'ils construisent  $\mathbf{D}(\boldsymbol{\theta})$  de manière à la rendre définie positive, ces algorithmes de Newton modifiés peuvent manipuler des problèmes où la fonction qu'ils faut minimiser n'est pas globalement convexe. Différents algorithmes choisissent  $\mathbf{D}(\boldsymbol{\theta})$  de différentes façons, dont certaines sont assez ingénieuses, et qu'il serait astucieux de faire exécuter par un ordinateur. Comme nous allons le voir cependant, il y a un moyen très simple et naturel de choisir  $\mathbf{D}(\boldsymbol{\theta})$ , pour des fonctions somme des carrés, fondé sur la régression de Gauss-Newton.

Dans de nombreux cas,  $\alpha^{(j)}$  est choisi de façon à minimiser  $Q(\boldsymbol{\theta}^{(j)} - \alpha^{(j)} (\mathbf{D}^{(j)})^{-1} \mathbf{g}^{(j)})$ , considérée comme une fonction unidimensionnelle en  $\alpha^{(j)}$ , de sorte que des points tels que  $\theta''$  dans le Figure 6.5 ne posent aucune

difficulté. Quelques algorithmes ne minimisent pas vraiment l'expression  $Q(\boldsymbol{\theta}^{(j)} - \alpha^{(j)}(\mathbf{D}^{(j)})^{-1}\mathbf{g}^{(j)})$  par rapport à  $\alpha^{(j)}$ , mais choisissent plutôt  $\alpha^{(j)}$  de façon à s'assurer que  $Q(\boldsymbol{\theta}^{(j+1)})$  est inférieur à  $Q(\boldsymbol{\theta}^{(j)})$ . Il est essentiel que cela soit vérifié si l'on veut être sûr que l'algorithme progressera vers la solution à chaque calcul. Les meilleurs algorithmes, qui ont été conçus pour économiser du temps de calcul à l'ordinateur, peuvent choisir  $\alpha$  de façon assez grossière lorsque le point de départ est loin de  $\hat{\boldsymbol{\theta}}$ , mais exécutent généralement une minimisation unidimensionnelle très fine lorsqu'ils sont proches de  $\hat{\boldsymbol{\theta}}$ .

Lorsque l'on essaie de minimiser la fonction somme-des-carrés  $SSR(\boldsymbol{\beta})$ , la régression de Gauss-Newton fournit un moyen très pratique d'approximer  $\mathbf{H}(\boldsymbol{\beta})$ . Les algorithmes qui adoptent cette procédure sont dits employer la **méthode de Gauss-Newton**. Dans la Section 5.4, nous avons vu que  $\mathbf{H}(\boldsymbol{\beta})$  est composée d'éléments dont la forme est

$$H_{il}(\boldsymbol{\beta}) = -2 \sum_{t=1}^n \left( (y_t - x_t(\boldsymbol{\beta})) \frac{\partial X_{ti}}{\partial \beta_l} - X_{ti}(\boldsymbol{\beta}) X_{tl}(\boldsymbol{\beta}) \right). \quad (6.39)$$

C'est l'équation (5.24) écrite avec une notation scalaire. Nous avons vu que lorsque le modèle est correct et que la matrice Hessienne est évaluée en la vraie valeur de  $\boldsymbol{\beta}$ , ceci est asymptotiquement équivalent à

$$2 \sum_{t=1}^n X_{ti}(\boldsymbol{\beta}) X_{tl}(\boldsymbol{\beta});$$

le résultat est (5.38). Par conséquent un choix naturel pour  $\mathbf{D}(\boldsymbol{\beta})$  dans un algorithme de minimisation du genre que nous avons décrit par (6.38) est

$$\mathbf{D}(\boldsymbol{\beta}) = 2\mathbf{X}^\top(\boldsymbol{\beta})\mathbf{X}(\boldsymbol{\beta}). \quad (6.40)$$

Le gradient de  $SSR(\boldsymbol{\beta})$  est

$$\mathbf{g}(\boldsymbol{\beta}) = -2\mathbf{X}^\top(\boldsymbol{\beta})(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})). \quad (6.41)$$

En introduisant (6.40) et (6.41) dans (6.38), on obtient

$$\begin{aligned} \boldsymbol{\beta}^{(j+1)} &= \boldsymbol{\beta}^{(j)} + \alpha^{(j)} (2(\mathbf{X}^{(j)})^\top \mathbf{X}^{(j)})^{-1} (2(\mathbf{X}^{(j)})^\top (\mathbf{y} - \mathbf{x}^{(j)})) \\ &= \boldsymbol{\beta}^{(j)} + \alpha^{(j)} ((\mathbf{X}^{(j)})^\top \mathbf{X}^{(j)})^{-1} (\mathbf{X}^{(j)})^\top (\mathbf{y} - \mathbf{x}^{(j)}) \\ &= \boldsymbol{\beta}^{(j)} + \alpha^{(j)} \mathbf{b}^{(j)}, \end{aligned}$$

où  $\mathbf{b}^{(j)}$  est l'estimation de  $\mathbf{b}$  dans la régression de Gauss-Newton avec  $\mathbf{x}(\boldsymbol{\beta})$  et  $\mathbf{X}(\boldsymbol{\beta})$  évaluées toutes deux en  $\boldsymbol{\beta}^{(j)}$ , et  $\mathbf{X}^{(j)} \equiv \mathbf{X}(\boldsymbol{\beta}^{(j)})$ .

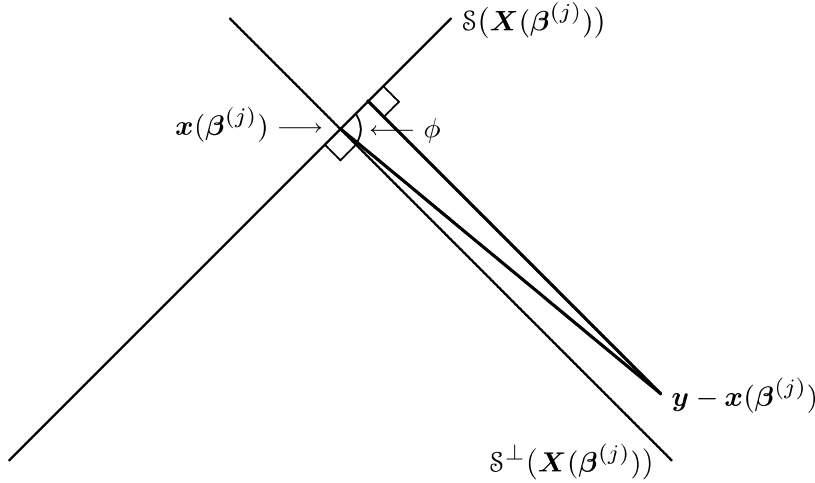
La relation entre les algorithmes de minimisation et la régression de Gauss-Newton est claire à présent. La GNR offre un moyen pratique et peu

coûteux d'obtenir une matrice qui approxime la matrice Hessienne de  $SSR(\beta)$  et qui soit toujours définie positive (sous condition que le modèle soit identifié en chacun des points où la GNR est exécutée). Elle offre encore plus que cela, puisque le vecteur des coefficients de la GNR est véritablement égal à  $-(D^{(j)})^{-1}g^{(j)}$ , qui est la direction dans laquelle l'algorithme exécutera les recherches à chaque étape. En combinant la GNR à un bon programme de recherche à une dimension (pour trouver  $\alpha$  à chaque étape), on arrive à obtenir un algorithme raisonnablement performant de recherche des estimations par moindres carrés non linéaires. Un tel algorithme représente une amélioration majeure de la "méthode de Gauss-Newton" d'origine, qui à l'image de la Méthode de Newton sous sa forme originelle initialise simplement  $\alpha$  à 1 à chaque étape.

Nous disons "raisonnablement performant" parce que des algorithmes plus évolués sont sans doute envisageables. La difficulté principale est que la matrice  $X^T(\beta)X(\beta)$  peut se révéler être quelquefois presque singulière même si le modèle est identifié relativement bien par les données lorsqu'il est évalué aux alentours  $\beta_0$ . Lorsque c'est le cas l'algorithme s'affole, parce que  $b$  ne se situe plus désormais dans le même espace à  $k$  dimensions que  $\beta$ , mais au contraire dans un espace dont la dimension est égale au rang de  $X^T(\beta)X(\beta)$ . On ne peut pas espérer trouver  $\hat{\beta}$  si on le cherche dans un espace dont le nombre de dimensions est trop faible, de sorte que quand cela survient un algorithme de Gauss-Newton non modifié entre dans une boucle infinie sans jamais réaliser de progrès. Les meilleurs algorithmes concernant les problèmes de moindres carrés vérifient si oui ou non ce phénomène survient, et choisissent pour  $D$  un objet qui se comporte mieux que  $X^T(\beta)X(\beta)$ , si cela est possible. Consulter les références citées précédemment.

La méthode de Gauss-Newton peut être aussi inefficace si  $2X^T(\beta)X(\beta)$  est une approximation médiocre de  $H(\beta)$ , autrement dit si le premier terme dans la somme de (6.39) est important. Cela peut arriver avec un modèle bien spécifié lorsque la taille de l'échantillon est faible, ou avec un modèle mal spécifié quelle que soit la taille de l'échantillon. Evidemment, l'inférence basée sur la théorie asymptotique sera peu fiable dans le premier cas, et sans espoir dans le second, ce qui fait du fonctionnement médiocre de la méthode de Gauss-Newton un élément de contrôle utile.

Aucune procédure de minimisation numérique ne trouve  $\hat{\theta}$  avec *exactitude*, et étant donnée les contraintes de l'arithmétique à virgule flottante sur les ordinateurs digitaux, il est souvent irréaliste d'espérer plus de six ou peut-être sept décimales de précision. A moins qu'ils ne soient avertis explicitement de s'arrêter, ces algorithmes itératifs poursuivent leur recherche, même si les modifications dans  $\beta$  et  $SSR(\beta)$  sont imputables à des erreurs d'arrondi de la part de l'ordinateur. Le choix de **règles d'arrêt** est par conséquent une partie importante de l'art de la minimisation non linéaire. De nombreux auteurs (voir Quandt (1983)) ont suggéré que la règle la plus naturelle pour arrêter



**Figure 6.6** La règle d'arrêt n'est pas pleinement satisfaite

un algorithme de Gauss-Newton est

$$\frac{(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)}))^{\top} \mathbf{X}^{(j)} (\mathbf{X}^{(j)\top} \mathbf{X}^{(j)})^{-1} \mathbf{X}^{(j)\top} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)}))}{(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)}))^{\top} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)})) / n} < \varepsilon, \quad (6.42)$$

où  $\varepsilon$  est un critère de convergence prédéterminé qui peut être modifié par l'utilisateur. Le membre de gauche (6.42) est  $n$  fois le  $R^2$  non centré de la régression de Gauss-Newton de  $\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)})$  sur  $\mathbf{X}^{(j)}$ , qui prend la même forme que la statistique de test dont nous avons parlé à la Section 6.4. Remarquons que, pourvu que  $\boldsymbol{\beta}^{(j)}$  soit proche de  $\boldsymbol{\beta}_0$ , l'expression (6.42) est approximativement égale à  $(SSR(\boldsymbol{\beta}^{(j)}) - SSR(\hat{\boldsymbol{\beta}})) / \sigma_0^2$ . Par là, cette règle d'arrêt indique à l'algorithme de s'arrêter lorsqu'une approximation de  $SSR(\boldsymbol{\beta})$  en  $\boldsymbol{\beta}^{(j)}$  assure que la distance entre  $SSR(\boldsymbol{\beta}^{(j)})$  et  $SSR(\hat{\boldsymbol{\beta}})$  est suffisamment faible relativement à la variance des aléas.

Une interprétation géométrique de cette règle d'arrêt est illustrée à la Figure 6.6. Le dénominateur du membre de gauche de (6.42) est

$$\frac{1}{n} \|\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)})\|^2, \quad (6.43)$$

qui est  $1/n$  fois le carré de la distance entre  $\mathbf{y}$  et  $\mathbf{x}(\boldsymbol{\beta}^{(j)})$ . Le numérateur est

$$\|\mathbf{P}_{\mathbf{X}^{(j)}}(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)}))\|^2, \quad (6.44)$$

qui est le carré de la longueur du vecteur résultant de la projection de  $\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)})$  sur  $S(\mathbf{X}^{(j)})$ . Le rapport de (6.44) par (6.43), ainsi que toutes les quantités qui s'interprètent comme des  $R^2$ , est le carré du cosinus d'un angle donné. Dans ce cas précis, l'angle dont il s'agit est celui formé par

$\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)})$  et sa projection sur  $\mathcal{S}(\mathbf{X}^{(j)})$ , que l'on a étiqueté  $\phi$  sur la figure. Lorsque la valeur de ce rapport est suffisamment faible,  $\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^{(j)})$  doit être presque orthogonal à  $\mathbf{X}^{(j)}$ , et les conditions du premier ordre sont également presque satisfaites. La règle d'arrêt (6.42) correspond à  $n$  fois  $\cos^2 \phi$ . Bien que le facteur  $n$  puisse être ignoré, il donne à la règle d'arrêt des propriétés appréciables. Ce facteur donne au critère que l'on compare à  $\varepsilon$  la forme d'une statistique de test  $nR^2$ , et il garantit que l'exactitude *numérique* des estimations est proportionnelle à  $n^{-1/2}$ , tout comme leur exactitude statistique. Toutefois, si la taille de l'échantillon était particulièrement importante, il faudrait alors veiller à ne pas utiliser une valeur trop faible pour  $\varepsilon$ , car autrement la règle d'arrêt nécessiterait un degré de précision numérique que n'atteindrait aucun ordinateur.

Cette discussion suggère que la règle d'arrêt (6.42) possède un attrait non négligeable. Gill, Murray, et Wright (1981) traitent un grand nombre de règles d'arrêt, qui souffrent toutes de certaines insuffisances (parmi elles le fait d'être sensibles à des variations d'échelle des paramètres) auxquelles (6.42) n'est pas sujette.<sup>3</sup> Parce qu'elle ne souffre pas de ces handicaps, mais aussi parce que le membre de gauche de l'inégalité se calcule aisément comme un produit dérivé de la GNR pour chaque  $\boldsymbol{\beta}^{(j)}$ , (6.42) semble être en réalité une règle d'arrêt très intéressante.

Bien évidemment, toute règle d'arrêt échouerait sans doute si  $\varepsilon$  était choisi de façon incorrecte. Lorsque  $\varepsilon$  est trop fort, l'algorithme peut arrêter les itérations trop tôt, alors que  $\boldsymbol{\beta}$  est encore assez éloigné de  $\hat{\boldsymbol{\beta}}$ . Dans le cas contraire, l'algorithme peut poursuivre l'itération alors que  $\boldsymbol{\beta}$  est très proche de  $\hat{\boldsymbol{\beta}}$ , les différences éventuelles pourront provenir des erreurs d'arrondi de sorte que l'algorithme peut ne jamais s'arrêter. Il est donc quelquefois enrichissant de tester des valeurs de  $\varepsilon$  pour connaître la sensibilité des résultats obtenus. Si la valeur affichée  $\hat{\boldsymbol{\beta}}$  varie dans des proportions notables lorsque  $\varepsilon$  est réduit, cela provient soit de la valeur initiale de  $\varepsilon$  qui était trop importante, soit de la puissance de l'algorithme qui tente de trouver un minimum précis. Les valeurs raisonnables de  $\varepsilon$  seraient, pour la règle d'arrêt que nous proposons, comprises entre  $10^{-4}$  et  $10^{-12}$ .

## 6.9 LECTURES COMPLÉMENTAIRES

Ainsi que nous l'avons noté, la régression de Gauss-Newton a été utilisée depuis de nombreuses années et en tant qu'élément clef de la méthode de Gauss-Newton qui possède en réalité de nombreux points communs avec les

<sup>3</sup> Étrangement, Gill, Murray, et Wright (1981) ne considèrent pas cette règle d'arrêt ou une quelconque de ses généralisations. C'est peut-être parce que (6.42) est une règle dont l'usage est immédiat pour la régression de Gauss-Newton, mais qui n'est pas évidente si l'on traite des problèmes plus généraux de minimisation.

algorithmes d'estimation par moindres carrés non linéaires. On peut trouver une discussion approfondie sur ce sujet chez Bard (1974). La méthode de Newton, ainsi que son nom le laisse supposer, est assez ancienne et l'idée que l'on peut approximer la matrice Hessienne par une matrice qui ne dépend que des dérivées premières date de Gauss (1809). Toutefois, comme l'estimation non linéaire était peu commode jusqu'à ce que des ordinateurs ne soient largement exploitables, la plupart des travaux sur ce sujet sont relativement récents. Des articles fondamentaux datant de la période postérieure au développement des calculateurs, consacrée à la méthode de Gauss-Newton sont ceux de Hartley (1961) et Marquardt (1963). L'article de Quandt (1983) fournit de nombreuses références, ainsi que Seber et Wild (1989, Chapitre 14).

Par contraste avec son usage ancien pour l'estimation, l'utilisation de la GNR dans les tests de spécification est assez récente. Le premier article dans la littérature économétrique semble dater de Durbin (1970), qui proposa ce qui se ramène à un cas particulier de la GNR lorsque l'on désire tester des modèles de régression linéaire appliqués à des cas d'autocorrélation avec des variables dépendantes retardées. Cependant, cette démarche a été traitée de façon plutôt superficielle puisque c'est à l'occasion du même article que Durbin proposa son fameux test  $h$ . Ce qui allait être connu sous le nom de "procédure alternative de Durbin", qui correspond effectivement à un cas particulier de la GNR, fut largement méconnue par les théoriciens pendant des années, et totalement ignorée par les praticiens. Mais nous verrons tout cela au Chapitre 10.

L'intérêt porté à la régression de Gauss-Newton en tant que procédure de construction de statistiques de test date de la fin des années 1970. Godfrey (1978a, 1978b) et Breusch (1978) généralisèrent la procédure alternative de Durbin et montrèrent comment on peut calculer un test LM pour l'autocorrélation à l'aide de la GNR. Une pléiade d'autres auteurs traitèrent d'autres cas particuliers, contribuant à améliorer la compréhension du cas général que nous avons abordé au cours de ce chapitre, et des tests qui s'y rattachent. Les articles de Breusch et Pagan (1980) et Engle (1982a) sont, à ce titre, remarquables. La grande partie de cette littérature suppose explicitement la normalité des erreurs, et élabore des tests comme des tests LM bâtis sur la structure de l'estimation par maximum de vraisemblance. Cela peut se révéler légèrement trompeur, car, comme nous l'avons vu, il n'y a aucune nécessité de supposer la normalité des erreurs, ni pour les estimations par moindres carrés non linéaires, ni pour les tests fondés sur la GNR, pour obtenir de résultats asymptotiquement valides. Des articles plus récents, parmi lesquels Pagan (1984a), Davidson et MacKinnon (1985a), et MacKinnon (1992), sont consacrés aux modèles de régression, et tentent d'unifier et de clarifier la littérature déjà existante. Nous aurons l'occasion de rencontrer de nombreuses régressions de Gauss-Newton, ainsi que des régressions artificielles connexes qui ont les mêmes propriétés, à travers tout le livre.

## TERMES ET CONCEPTS

algorithmes d'optimisation numérique	nombre condition (d'une matrice)
algorithmes des moindres carrés non linéaires	règles d'arrêt
colinéarité (dans les régressions artificielles) et identifiabilité	régression artificielle
estimateurs efficaces en une étape	régression augmentée
estimation de la matrice de covariance	régression de Gauss-Newton (GNR)
robuste à l'hétéroscédasticité (HCCME)	test d'erreur de spécification de la régression (RESET)
identification et colinéarité	tests à variable additionnelle
méthode de Gauss-Newton	tests $C(\alpha)$
méthode de Newton	tests diagnostiques
modèle insuffisamment identifié	tests $nR^2$ et tests en $F$ fondés sur la GNR
	tests pseudo-Wald



# Chapitre 7

## Les Variables Instrumentales

### 7.1 INTRODUCTION

La seule technique d'estimation que nous ayons considérée jusqu'à présent est celle des moindres carrés ordinaires et non linéaires. Bien que les moindres carrés comportent de nombreux avantages, ils possèdent aussi de nombreux inconvénients. L'un des principaux inconvénients est qu'ils ne donnent des estimations convergentes que si les aléas sont asymptotiquement orthogonaux aux régresseurs, ou dans le cas non linéaire, aux dérivées de la fonction de régression. Considérons, pour simplifier, le modèle de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (7.01)$$

où  $\mathbf{X}$  est une matrice de dimension  $n \times k$  des variables explicatives. Les résultats sont les mêmes si la fonction de régression est linéaire ou non linéaire, et nous traiterons pour simplifier le cas linéaire. Quand les données sont générées par le DGP

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I}), \quad (7.02)$$

nous avons vu que l'estimation OLS est

$$\hat{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}. \quad (7.03)$$

Il est évident que si le vecteur  $\hat{\boldsymbol{\beta}}$  doit converger vers  $\boldsymbol{\beta}_0$ , la condition

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{u}) = \mathbf{0}$$

doit être vérifiée. Si l'estimation  $\hat{\boldsymbol{\beta}}$  n'est pas biaisée, la condition plus forte que  $E(\mathbf{X}^\top \mathbf{u}) = \mathbf{0}$  doit également être vérifiée. Ces conditions nécessaires ne sont pas directement vérifiables, puisque la propriété d'orthogonalité des moindres carrés garantit qu'il est indifférent que  $\mathbf{u}$  soit corrélé ou non avec  $\mathbf{X}$ , les résidus  $\hat{\mathbf{u}}$  étant orthogonaux à  $\mathbf{X}$ . Cela signifie que, peu importe que les estimations par moindres carrés puissent être biaisées et non convergentes, les résidus des moindres carrés ne fourniront pas de preuves s'il y a un problème.

Supposons que  $\text{plim}(n^{-1}\mathbf{X}^\top \mathbf{u}) = \mathbf{w}$ , un vecteur non nul. Alors, à partir de (7.03) il est clair que  $\text{plim}(\hat{\beta}) \neq \beta_0$ . De plus, la limite en probabilité de  $n^{-1}$  fois la somme des carrés des résidus sera

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{u}^\top \mathbf{M}_X \mathbf{u}) = \sigma_0^2 - \mathbf{w}^\top \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{w}.$$

Si  $\mathbf{u}$  était non corrélé asymptotiquement avec  $\mathbf{X}$ , cette quantité serait alors  $\sigma_0^2$ . Au lieu de cela, elle est *plus petite* que  $\sigma_0^2$ . Par conséquent, l'utilisation des moindres carrés provoque un ajustement trop parfait du modèle. Parce que les moindres carrés minimisent la distance entre  $\mathbf{y}$  et  $\mathcal{S}(\mathbf{X})$ , une partie de la variation de  $\mathbf{y}$  qui est en vérité imputable à la variation de l'aléa  $\mathbf{u}$ , a été attribuée à tort à une variation des régresseurs.

Malheureusement en économétrie, dans plusieurs situations, les aléas ne peuvent être supposés orthogonaux à la matrice  $\mathbf{X}$ . Nous discuterons précisément de deux cas dans les Sections 7.2 et 7.3, celui de l'erreur dans les variables et celui du biais dans les équations simultanées. La méthode la plus générale pour traiter de tels cas est celle des **variables instrumentales** ou **IV**. Cette méthode, proposée à l'origine par Reiersøl (1941) et développée plus tard par Durbin (1954) et Sargan (1958), parmi tant d'autres auteurs, est très puissante et générale. De nombreuses variantes de cette méthode apparaissent dans beaucoup de domaines de l'économétrie. Cela inclut: **doubles moindres carrés** (Section 7.5), les **triples moindres carrés** (Chapitre 18), et la **méthode généralisée des moments** (Chapitre 17).

Le plan de ce chapitre est le suivant. Dans la prochaine section, nous discuterons du problème commun des erreurs dans les variables pour lequel, à l'origine, la méthode des variables instrumentales a été proposée comme une solution. Alors, dans la Section 7.3, nous ferons une introduction au modèle linéaire des équations simultanées et nous montrerons que l'estimation par OLS est biaisée lorsqu'elle est appliquée à une des équations du modèle. Dans la Section 7.4, nous introduirons la méthode des variables instrumentales dans le contexte d'une équation de régression linéaire et nous discuterons ses nombreuses propriétés. Dans la section suivante, nous discuterons de l'autre nom pour l'estimateur IV des paramètres du modèle de régression linéaire, à savoir les doubles moindres carrés. Dans la Section 7.6, nous montrerons comment la méthode IV peut être utilisée pour estimer des modèles de régression non linéaire. Dans la Section 7.7, nous généraliserons la régression de Gauss-Newton au cas IV et nous verrons comment tester les hypothèses des coefficients des modèles de régression lorsqu'ils ont été estimés par IV. Dans la Section 7.8, nous aborderons le problème de l'identification dans les modèles de régression estimés par IV. Enfin, dans la Section 7.9, nous considérerons une classe de tests appelés tests de Durbin-Wu-Hausman, qui peuvent être utilisés pour décider s'il est nécessaire ou non d'employer les variables instrumentales.

## 7.2 LES ERREURS DANS LES VARIABLES

Presque toutes les variables économiques sont mesurées avec erreur. Cela se produit dans une mesure plus ou moins grande pour toutes les variables macroéconomiques temporelles, et cela est particulièrement vrai avec des données d'enquêtes et de nombreux autres ensembles de données en coupe transversale. Malheureusement, les erreurs dans les variables explicatives ont des conséquences statistiques néfastes, puisque celles mesurées avec erreur sont nécessairement corrélées avec les aléas. Quand cela se produit, le problème est dit être un problème d'**erreurs dans les variables**. Nous illustrons le problème des erreurs dans les variables à partir d'un exemple simple.

Supposons, pour simplifier, que le DGP soit

$$\mathbf{y} = \alpha_0 + \beta_0 \mathbf{x} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_0^2 \mathbf{I}), \quad (7.04)$$

où  $\mathbf{x}$  est un vecteur qui est observé avec erreur. Nous observons réellement  $\mathbf{x}^*$ , qui est relié à  $\mathbf{x}$  par

$$\mathbf{x}^* = \mathbf{x} + \mathbf{v}, \quad \mathbf{v} \sim \text{IID}(\mathbf{0}, \omega^2 \mathbf{I}).$$

Le vecteur  $\mathbf{v}$  mesure les erreurs, qui sont supposées, sans doute de façon peu réaliste, être i.i.d. et indépendantes de  $\mathbf{x}$  et de  $\mathbf{u}$ . En remplaçant  $\mathbf{x}$  par  $\mathbf{x}^* - \mathbf{v}$  dans (7.04), le DGP devient

$$\mathbf{y} = \alpha_0 + \beta_0 \mathbf{x}^* - \beta_0 \mathbf{v} + \mathbf{u}.$$

Par conséquent, l'équation effectivement estimée est

$$\mathbf{y} = \alpha + \beta \mathbf{x}^* + \mathbf{u}^*, \quad (7.05)$$

où  $\mathbf{u}^* \equiv \mathbf{u} - \beta_0 \mathbf{v}$ . Il est clair que  $\mathbf{u}^*$  n'est pas indépendant de  $\mathbf{x}^*$ . En fait,

$$E(\mathbf{x}^{*\top} \mathbf{u}^*) = E((\mathbf{x} + \mathbf{v})^\top (\mathbf{u} - \beta_0 \mathbf{v})) = -\beta_0 E(\mathbf{v}^\top \mathbf{v}) = -n\beta_0 \omega^2,$$

où, comme d'habitude,  $n$  est la taille de l'échantillon. Si nous supposons concrètement que  $\beta_0 > 0$ , l'aléa  $\mathbf{u}^*$  est corrélé négativement avec le régresseur  $\mathbf{x}^*$ . Cette corrélation négative signifie que l'estimation par moindres carrés de  $\beta$  sera biaisée et non convergente, comme celle de  $\alpha$ , à moins que  $\mathbf{x}^*$  ne soit de moyenne nulle. Notons que la non convergence de  $\hat{\beta}$  constitue un problème seulement si nous nous intéressons au paramètre  $\beta$ . Si, au contraire, nous avons cherché l'espérance de  $\mathbf{y}$  conditionnelle à  $\mathbf{x}^*$ , nous aurions pu estimer l'équation (7.05) par moindres carrés.

Il existe plusieurs moyens de traiter le problème des erreurs dans les variables, la méthode des variables instrumentales en est précisément un. Dans l'exemple précédent, il est clair que si nous connaissions  $\omega^2$ , nous pourrions dire quelque chose à propos du biais de  $\hat{\beta}$  et de ce fait, trouver une meilleure estimation. Cette observation a conduit vers différentes approches alternatives du problème des erreurs dans les variables. Voir parmi entre autres Frisch (1934), Klepper et Leamer (1984), Hausman et Watson (1985), et Leamer (1987).

### 7.3 LES EQUATIONS SIMULTANÉES

La raison la plus souvent évoquée dans les travaux d'économétrie appliquée pour que les variables explicatives soient corrélées avec les aléas est que les variables explicatives sont déterminées de manière endogène plutôt que d'être exogènes ou prédéterminées. Une variable dite **prédéterminée** au temps  $t$  est une variable qui a été déterminée, peut-être de manière endogène, à une période antérieure. L'exemple le plus simple est celui d'une variable dépendante retardée. Une discussion détaillée sur l'exogénéité et la prédétermination est proposée à la Section 18.2. Les modèles dans lesquels deux ou plusieurs variables endogènes sont déterminées simultanément, sont appelés **modèles à équations simultanées**. En effet, depuis plusieurs années, le modèle linéaire à équations simultanées a été le centre d'intérêt de la théorie économétrique, et la littérature traitant de la manière d'estimer ces modèles demeure assez vaste. Nous consacrerons le Chapitre 18 à la discussion de ce sujet. Dans cette section, nous discuterons seulement d'un exemple très simple, celui d'un modèle linéaire à deux équations de détermination du prix et de la quantité sur un marché concurrentiel. Cet exemple illustre plusieurs résultats et concepts de base dans l'analyse des modèles à équations simultanées. En particulier, il est clair qu'il y aura généralement corrélation entre les aléas et les variables endogènes placées à droite de l'égalité. De plus, le modèle d'offre-demande a constitué un des premiers centres d'intérêt pour le développement des méthodes de traitement des modèles à équations simultanées; consulter Goldberger (1972).

Le modèle que nous traitons est le suivant:

$$Q_t^d = \alpha P_t + \mathbf{Z}_t^d \boldsymbol{\beta} + u_t^d \quad (7.06)$$

$$Q_t^s = \gamma P_t + \mathbf{Z}_t^s \boldsymbol{\delta} + u_t^s, \quad (7.07)$$

où  $Q_t^d$  correspond à la quantité demandée pour l'observation  $t$ ,  $Q_t^s$  correspond à la quantité offerte,  $P_t$  est le prix,  $\mathbf{Z}_t^d$  désigne un vecteur de variables exogènes et/ou prédéterminées dans la fonction de demande, et  $\mathbf{Z}_t^s$  représente un vecteur de variables exogènes et/ou prédéterminées dans la fonction d'offre. Le prix et la quantité pourraient être exprimés en logarithmes plutôt qu'en niveaux, puisqu'une spécification log-linéaire serait plus acceptable pour les fonctions de demande et d'offre. Si nos données proviennent d'un marché concurrentiel toujours en équilibre (hypothèse qui ne serait pas toujours plausible dans chaque cas), nous savons que

$$Q_t^d = Q_t^s = Q_t,$$

où  $Q_t$  est la quantité effectivement vendue. Ainsi le prix  $P_t$  est supposé être déterminé de manière endogène par l'égalisation de (7.06) et (7.07). Il est évident que le prix et la quantité sont déterminés simultanément dans ce modèle.

A présent, nous voulons écrire la **forme structurelle** de ce modèle. Nous devons donc pour cela remplacer  $Q_t^d$  et  $Q_t^s$  par  $Q_t$ , et récrire les fonctions de demande et d'offre (7.06) et (7.07) en terme des variables observées  $P_t$  et  $Q_t$ . Il existe plusieurs moyens équivalents de procéder. La fonction de demande, l'équation (7.06), peut être réécrite sous l'une des deux formes :

$$Q_t = \alpha P_t + \mathbf{Z}_t^d \boldsymbol{\beta} + u_t^d, \text{ ou} \quad (7.08a)$$

$$P_t = \alpha^* Q_t + \mathbf{Z}_t^d \boldsymbol{\beta}^* + u_t^{d*}. \quad (7.08b)$$

De la même manière, la fonction d'offre, l'équation (7.07), peut être réécrite sous l'une des deux formes :

$$Q_t = \gamma P_t + \mathbf{Z}_t^s \boldsymbol{\delta} + u_t^s, \text{ ou} \quad (7.09a)$$

$$P_t = \gamma^* Q_t + \mathbf{Z}_t^s \boldsymbol{\delta}^* + u_t^{s*}. \quad (7.09b)$$

Les quantités avec une astérisque dans les équations *b* sont reliées, de manière évidente, à celles sans astérisque dans les équations *a*. Par exemple, les paramètres et les aléas de (7.08b) sont reliés avec ceux de (7.08a) de la manière suivante:

$$\alpha^* = \alpha^{-1}; \quad \boldsymbol{\beta}^* = -\alpha^{-1} \boldsymbol{\beta}; \quad u_t^{d*} = -\alpha^{-1} u_t^d.$$

Lorsque l'on écrit le modèle en entier, il est possible de combiner soit (7.08a) avec (7.09a) ou (7.09b), soit (7.08b) avec (7.09a) ou (7.09b). Nous disposons donc de *quatre* manières différentes d'écrire ce système d'équations, chaque manière étant aussi valable qu'une autre. Il est conventionnel d'écrire les modèles à équations simultanées de manière à ce que chaque variable endogène apparaisse du côté gauche d'une seule équation. Mais il n'y a rien de sacro-saint à propos de cette convention. En effet, du point de vue de la théorie économique, il est probablement plus naturel de combiner (7.08a) avec (7.09a), en mettant la quantité dans le membre de gauche des équations de demande et d'offre.

Nous venons juste de voir que la **normalisation** (c'est-à-dire savoir quelle variable endogène associer à un coefficient unitaire et placer dans le membre de gauche de chaque équation) est nécessaire quand nous traitons un système à équations simultanées. Il n'existe pas une seule manière d'écrire le système parce qu'il y a deux ou plusieurs variables endogènes. Par conséquent, contrairement à ce que les développements sur ce sujet ont suggéré, et il n'existe pas de forme structurelle unique pour un modèle linéaire à équations simultanées. Il existe autant de formes structurelles que de manières de normaliser le système d'équations.

Les formes structurelles d'un modèle à équations simultanées doivent être contrastées avec les **formes réduites** qui sont de deux variétés. La **forme réduite contrainte**, ou **RRF**, nécessite de récrire le modèle pour que chaque

variable endogène n'apparaisse qu'une seule fois. Pour y parvenir, nous commençons par écrire la forme structurelle composée de (7.08a) et (7.09a):

$$Q_t - \alpha P_t = \mathbf{Z}_t^d \boldsymbol{\beta} + u_t^d$$

$$Q_t - \gamma P_t = \mathbf{Z}_t^s \boldsymbol{\delta} + u_t^s.$$

Ces deux équations peuvent se récrire en utilisant une notation matricielle

$$\begin{bmatrix} 1 & -\alpha \\ 1 & -\gamma \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_t^d \boldsymbol{\beta} \\ \mathbf{Z}_t^s \boldsymbol{\delta} \end{bmatrix} + \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix}.$$

En résolvant ce système pour  $Q_t$  et  $P_t$ , nous obtenons la forme réduite contrainte :

$$\begin{bmatrix} Q_t \\ P_t \end{bmatrix} = \begin{bmatrix} 1 & -\alpha \\ 1 & -\gamma \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z}_t^d \boldsymbol{\beta} \\ \mathbf{Z}_t^s \boldsymbol{\delta} \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ 1 & -\gamma \end{bmatrix}^{-1} \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix},$$

système qui peut s'écrire de manière plus explicite comme

$$Q_t = \frac{1}{\alpha - \gamma} (\alpha \mathbf{Z}_t^s \boldsymbol{\delta} - \gamma \mathbf{Z}_t^d \boldsymbol{\beta}) + v_t^1 \quad (7.10)$$

$$P_t = \frac{1}{\alpha - \gamma} (\mathbf{Z}_t^s \boldsymbol{\delta} - \mathbf{Z}_t^d \boldsymbol{\beta}) + v_t^2, \quad (7.11)$$

où les aléas  $v_t^1$  et  $v_t^2$  sont des combinaisons linéaires des aléas originaux  $u_t^d$  et  $u_t^s$ .

Observons que les équations de la RRF, (7.10) et (7.11), sont non linéaires en leurs paramètres mais linéaires en leurs variables  $\mathbf{Z}_t^d$  et  $\mathbf{Z}_t^s$ . En fait, ce sont de simples versions contraintes de la **forme réduite non contrainte**, où **URF**,

$$Q_t = \mathbf{Z}_t \boldsymbol{\pi}_1 + v_t^1 \quad (7.12)$$

$$P_t = \mathbf{Z}_t \boldsymbol{\pi}_2 + v_t^2, \quad (7.13)$$

où  $\mathbf{Z}_t$  désigne un vecteur composé de toutes les variables qui apparaissent soit en  $\mathbf{Z}_t^d$  soit en  $\mathbf{Z}_t^s$ , et  $\boldsymbol{\pi}_1$  et  $\boldsymbol{\pi}_2$  sont des vecteurs de paramètres. Les deux équations de la URF peuvent être évidemment estimées, de façon convergente, par OLS puisque seules les variables exogènes ou prédéterminées apparaissent du côté droit de l'équation. La RRF serait plus difficile à estimer puisque cela implique des contraintes non linéaires à travers les équations. En fait, estimer la RRF est équivalent à estimer la forme structurelle sur laquelle elle est basée, comme nous le verrons dans le Chapitre 18.

Si nous nous contentions simplement d'estimer la URF, nous pourrions nous arrêter à ce point, puisque les estimations par OLS de (7.12) et (7.13)

seraient, à l'évidence, convergentes.<sup>1</sup> Cependant, les économistes préfèrent souvent estimer une forme structurelle d'un modèle à équations simultanées, soit parce que les paramètres de ces formes structurelles suscitent de l'intérêt, soit parce qu'imposer des contraintes croisées implicites dans la forme structurelle peut conduire, en grande partie, à une efficacité augmentée. Par conséquent, il semble intéressant de se demander ce qui se passe si nous appliquons les OLS à une des équations de l'une de ces formes structurelles. Considérons l'équation (7.08a). Les estimations par OLS de  $\alpha$  et  $\beta$  sont

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^\top \mathbf{P} & \mathbf{P}^\top \mathbf{Z}_d \\ \mathbf{Z}_d^\top \mathbf{P} & \mathbf{Z}_d^\top \mathbf{Z}_d \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{P}^\top \mathbf{Q} \\ \mathbf{Z}_d^\top \mathbf{Q} \end{bmatrix},$$

où  $\mathbf{P}$  et  $\mathbf{Q}$  désignent les vecteurs des observations sur  $P_t$  et  $Q_t$ , et  $\mathbf{Z}_d$  désigne la matrice des observations sur  $\mathbf{Z}_t^d$ . Si nous supposons que le modèle est correctement spécifié et que nous remplaçons  $\mathbf{Q}$  par  $\alpha_0 \mathbf{P} + \mathbf{Z}_d \beta_0 + \mathbf{u}_d$ , nous obtenons

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \mathbf{P}^\top \mathbf{P} & \mathbf{P}^\top \mathbf{Z}_d \\ \mathbf{Z}_d^\top \mathbf{P} & \mathbf{Z}_d^\top \mathbf{Z}_d \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{P}^\top \mathbf{u}_d \\ \mathbf{Z}_d^\top \mathbf{u}_d \end{bmatrix}. \quad (7.14)$$

Il est évident que ces estimations seront biaisées et non convergentes. Elles ne peuvent pas être sans biais, puisque la variable endogène  $P_t$  apparaît à droite dans l'équation. Elles seront non convergentes parce que

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{P}^\top \mathbf{u}_d) \neq 0,$$

puisque le prix d'équilibre dépend, en partie, de l'aléa dans l'équation de demande. Par conséquent, l'hypothèse standard d'indépendance entre les aléas et les régresseurs est enfreinte dans ce système (et dans tout système) d'équations simultanées. Par conséquent, si nous essayons de prendre la limite en probabilité du côté droit du système (7.14), nous trouverons que le second terme n'est pas égal à zéro. Il en résulte que les estimations  $\hat{\alpha}$  et  $\hat{\beta}$  ne seront pas convergentes.

Les résultats de ce simple exemple sont vrais en toute généralité. Étant donné qu'elles sont déterminées simultanément, toutes les variables endogènes dans un système d'équations simultanées dépendent généralement de l'aléa dans toutes les équations. Par conséquent, excepté dans un très petit nombre de cas, les variables endogènes placées à droite dans l'équation structurelle d'un système seront toujours corrélées avec les aléas. Donc, l'application des OLS à une telle équation donnerait des estimations biaisées et non convergentes.

<sup>1</sup> Il peut sembler que l'estimation OLS de la URF ne soit pas efficace, parce que les aléas de (7.12) et (7.13) seront corrélés. Cependant, comme nous le verrons dans le Chapitre 9, cette corrélation ne peut pas être exploitée pour donner des estimations plus efficaces parce que les régresseurs dans les deux équations sont les mêmes.

Nous venons de voir deux situations importantes dans lesquelles les variables explicatives seront corrélées avec l'aléa des équations de régression et nous sommes capables d'aborder le thème principal de ce chapitre, qui n'est autre que la méthode des variables instrumentales. Cette méthode peut être utilisée lorsque les aléas sont corrélés avec une ou plusieurs variables explicatives, sans se soucier de l'origine de cette corrélation. Cette méthode est donc simple, générale et puissante.

## 7.4 LES VARIABLES INSTRUMENTALES: LE CAS LINÉAIRE

L'élément fondamental de toute procédure IV est une matrice de **variables instrumentales** (ou simplement **instruments**). Nous appellerons cette matrice  $\mathbf{W}$  et nous spécifierons qu'elle est de dimension  $n \times l$ . Les colonnes de  $\mathbf{W}$  sont constituées de variables exogènes ou/et prédéterminées dont nous savons qu'elles sont indépendantes des aléas  $\mathbf{u}$  (ou du moins nous le supposons). Dans le contexte d'un modèle à équations simultanées, un choix naturel pour  $\mathbf{W}$  sera la matrice des variables exogènes et prédéterminées du modèle. Il doit y avoir au moins autant d'instruments qu'il y existe de variables explicatives dans l'équation qui doit être estimée. Par conséquent, si l'équation à estimer est le modèle de régression linéaire (7.01), avec  $\mathbf{X}$  composé de  $k$  colonnes, il est nécessaire que  $l \geq k$ . Il s'agit d'une condition d'identification; voir la Section 7.8 pour davantage de développements sur les conditions d'identification des modèles estimés par IV. Certaines variables explicatives peuvent apparaître parmi les instruments. En effet, comme nous le verrons plus loin, si nous voulions obtenir des estimations asymptotiquement efficaces, certaines colonnes de  $\mathbf{X}$  dont nous savons qu'elles sont exogènes et prédéterminées, devraient être incluses dans  $\mathbf{W}$ .

Intuitivement, la procédure des variables instrumentales est la suivante. Les moindres carrés qui minimisent la distance entre  $\mathbf{y}$  et  $\mathcal{S}(\mathbf{X})$  nous conduisent à des estimations non convergentes parce que  $\mathbf{u}$  est corrélé avec  $\mathbf{X}$ . L'espace à  $n$  dimensions où  $\mathbf{y}$  est un point, peut être divisé en deux sous-espaces orthogonaux  $\mathcal{S}(\mathbf{W})$  et  $\mathcal{S}^\perp(\mathbf{W})$ . Les variables instrumentales minimisent seulement la partie de la distance entre  $\mathbf{y}$  et  $\mathcal{S}(\mathbf{X})$  qui se trouve dans  $\mathcal{S}(\mathbf{W})$ . À condition que  $\mathbf{u}$  soit indépendant de  $\mathbf{W}$ , toute corrélation entre  $\mathbf{u}$  et  $\mathbf{X}$  doit se trouver, asymptotiquement, dans  $\mathcal{S}^\perp(\mathbf{W})$ . En grand échantillon, restreindre la minimisation à  $\mathcal{S}(\mathbf{W})$  évite donc les conséquences d'une corrélation entre  $\mathbf{u}$  et  $\mathbf{X}$ .

De manière plus formelle, lorsque nous appliquons les OLS au modèle (7.01), nous minimisons la somme des carrés des résidus

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

relativement à  $\boldsymbol{\beta}$ . Par contraste, lorsque nous appliquons la procédure IV au même modèle, nous minimisons la **fonction critère**

$$\|\mathbf{P}_W(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{P}_W (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (7.15)$$



où  $\mathbf{P}_W$  est la matrice qui projette orthogonalement sur  $\mathcal{S}(\mathbf{W})$ . Les conditions du premier ordre qui caractérisent une solution au problème de minimisation sont

$$\mathbf{X}^\top \mathbf{P}_W (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{0}, \quad (7.16)$$

où  $\tilde{\boldsymbol{\beta}}$  désigne le vecteur des estimations par IV. Par conséquent, nous voyons que les résidus IV  $\tilde{\mathbf{u}}$  doivent être orthogonaux à la projection des colonnes de  $\mathbf{X}$  sur  $\mathcal{S}(\mathbf{W})$ . Cela contraste avec la situation des OLS où les résidus étaient simplement orthogonaux à  $\mathcal{S}(\mathbf{X})$ . En résolvant (7.16) pour  $\tilde{\boldsymbol{\beta}}$ , nous obtenons

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y}. \quad (7.17)$$

Dans ce cas, nous avons supposé que la matrice  $\mathbf{X}^\top \mathbf{P}_W \mathbf{X}$  était de plein rang, ce qui est une condition nécessaire pour que  $\tilde{\boldsymbol{\beta}}$  soit identifié.

Il est facile de montrer que l'estimateur IV  $\tilde{\boldsymbol{\beta}}$  est convergent si les données sont générées effectivement par le DGP (7.02) et si certaines hypothèses sont satisfaites. Ces hypothèses sont

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{W}^\top \mathbf{u}) = \lim_{n \rightarrow \infty} (n^{-1} E(\mathbf{W}^\top \mathbf{u})) = \mathbf{0}, \quad (7.18a)$$

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{W}^\top \mathbf{W}) = \lim_{n \rightarrow \infty} (n^{-1} E(\mathbf{W}^\top \mathbf{W})) \text{ existe, est finie} \quad (7.18b)$$

et est définie positive, et

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{W}) \text{ existe, est finie et de plein rang } k. \quad (7.18c)$$

L'hypothèse la plus critiquable est (7.18a), qui n'est malheureusement pas vérifiable entièrement mais qui peut être testée sous certaines conditions; voir Section 7.9.

En remplaçant (7.02) dans (7.17), nous voyons que

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W (\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}) \\ &= \boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{u}. \end{aligned} \quad (7.19)$$

Premièrement, nous observons que  $\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{X})$  égale

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{W}) \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{W}^\top \mathbf{W})^{-1} \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{W}^\top \mathbf{X}).$$

Par conséquent, il en résulte à partir de (7.18) que  $\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{X})$  existe, qu'elle est finie et définie positive. D'où

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} (\tilde{\boldsymbol{\beta}}) &= \boldsymbol{\beta}_0 + \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \\ &\quad \times \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{W}) \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{W}^\top \mathbf{W})^{-1} \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{W}^\top \mathbf{u}). \end{aligned}$$

Les hypothèses (7.18) impliquent à présent que le second terme soit, ici, égal à zéro; l'hypothèse critique étant (7.18a). Par conséquent, nous en concluons que l'estimateur  $\tilde{\beta}$  converge vers  $\beta_0$ .

Bien que l'estimateur IV soit convergent, nous remarquons qu'il n'est pas sans biais. Parce que les colonnes de  $\mathbf{X}$ , et probablement certaines colonnes de  $\mathbf{W}$ , sont stochastiques, il est clair que

$$E\left((\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{u}\right) \neq \mathbf{0} \quad (7.20)$$

même si nous supposons que  $E(\mathbf{W}^\top \mathbf{u}) = \mathbf{0}$ . Nous verrons par la suite que l'espérance dans (7.20) peut même ne pas exister dans certains cas. Lorsqu'elles existent les espérances des estimateurs IV sont généralement biaisées. Nous aurons beaucoup à dire sur ce sujet dans la prochaine section.

Si nous retenons l'hypothèse supplémentaire que  $n^{-1/2} \mathbf{W}^\top \mathbf{u}$  obéit à un Théorème de la Limite Centrale, l'estimateur IV sera asymptotiquement distribué suivant une loi normale avec une matrice de covariance particulière. Dans le cas où les instruments  $\mathbf{W}$  ne sont pas stochastiques, cette hypothèse suit immédiatement l'hypothèse (7.02), concernant la distribution de  $\mathbf{u}$ . À partir de l'expression de droite de (7.19), nous pouvons déduire que

$$n^{1/2}(\tilde{\beta} - \beta_0) = (n^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} n^{-1/2} \mathbf{X}^\top \mathbf{P}_W \mathbf{u}. \quad (7.21)$$

Le facteur  $n^{-1/2} \mathbf{X}^\top \mathbf{P}_W \mathbf{u}$  peut être récrit comme

$$n^{-1/2} \mathbf{X}^\top \mathbf{P}_W \mathbf{u} = (n^{-1} \mathbf{X}^\top \mathbf{W})(n^{-1} \mathbf{W}^\top \mathbf{W})^{-1} n^{-1/2} \mathbf{W}^\top \mathbf{u}, \quad (7.22)$$

et, puisque les facteurs du côté droit de cette équation autres que  $n^{-1/2} \mathbf{W}^\top \mathbf{u}$  ont grâce aux hypothèses (7.18) des limites en probabilité bien définies, il en résulte que (7.22) est distribué asymptotiquement selon une normale dont la matrice de covariance est

$$\sigma_0^2 \operatorname{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{X}).$$

Par conséquent, nous en concluons à partir de (7.21) que

$$n^{1/2}(\tilde{\beta} - \beta_0) \overset{a}{\sim} N\left(\mathbf{0}, \sigma_0^2 \operatorname{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1}\right), \quad (7.23)$$

où, comme d'habitude,  $\overset{a}{\sim}$  signifie "est distribué asymptotiquement suivant".

En pratique, nous nous intéressons à la matrice de covariance de  $\tilde{\beta} - \beta_0$  plutôt qu'à celle de  $n^{1/2}(\tilde{\beta} - \beta_0)$ , et cela sans connaître  $\sigma_0$ . Nous pourrions estimer  $\sigma^2$  par

$$\tilde{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\tilde{\beta})^\top(\mathbf{y} - \mathbf{X}\tilde{\beta}).$$

Dans ce cas, il serait bien entendu possible de diviser par  $n - k$  plutôt que par  $n$ . Mais cela n'est pas nécessairement une bonne idée. Puisque la SSR *n'est pas* pour l'estimation IV la valeur de la fonction objectif (ce qui contraste avec les OLS), cette espérance n'est pas nécessairement plus petite que  $n\sigma_0^2$  et n'est certainement pas égale à  $(n - k)\sigma_0^2$ . Asymptotiquement, il n'existe aucune différence entre la division par  $n$  et celle par  $n - k$ . Quelle que soit la façon dont nous définissons  $\tilde{\sigma}$ , nous estimerons la matrice de covariance de  $\tilde{\beta} - \beta_0$  par

$$\tilde{\mathbf{V}}(\tilde{\beta}) = \tilde{\sigma}^2(\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1}. \quad (7.24)$$

L'estimateur IV  $\tilde{\beta}$  dont nous avons discuté, est un **estimateur IV généralisé**. Il peut être comparé avec l'**estimateur IV simple** qui est traité dans beaucoup d'articles élémentaires de statistique et d'économétrie, et qui fut développé en premier. Pour l'estimateur IV simple, nous associons à chaque variable explicative un unique instrument, qui peut être la variable elle-même si elle est supposée être non corrélée à  $\mathbf{u}$ . Par conséquent, la matrice  $\mathbf{W}$  est de même dimension,  $n \times k$ , que la matrice  $\mathbf{X}$ . Dans ce cas particulier, l'estimateur IV généralisé (7.17) se simplifie en grande partie:

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{y} \\ &= (\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{W} (\mathbf{X}^\top \mathbf{W})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{y} \\ &= (\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{y} \\ &= (\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{y}. \end{aligned} \quad (7.25)$$

La clef du résultat est que

$$(\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X})^{-1} = (\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{W} (\mathbf{X}^\top \mathbf{W})^{-1},$$

qui dépend du fait que la matrice  $\mathbf{W}^\top \mathbf{X}$  est carrée et de plein rang. La dernière ligne de (7.25) est la formule de l'estimateur IV simple qui apparaît dans de nombreux ouvrages. Mais nous ne discuterons pas plus de cet estimateur. Nous le rencontrerons à nouveau lorsque nous discuterons de la méthode généralisée des moments dans le Chapitre 17.

En pratique, le problème le plus important dans l'utilisation des IV réside dans le choix de la matrice des instruments  $\mathbf{W}$ . Bien que chaque ensemble valable d'instruments produise des estimations convergentes, différents choix

donneront des estimations différentes pour un échantillon fini. Lorsque nous utilisons des séries temporelles, il est naturel d'employer des variables retardées, incluant des valeurs retardées de la variable dépendante comme instruments. Mais nous ne connaissons pas pour autant le nombre des retards à utiliser. Lorsque nous estimons une équation à partir d'un modèle à équations simultanées, un ensemble naturel d'instruments est l'ensemble de toutes les variables exogènes et prédéterminées dans le modèle. Cependant, il peut y en avoir un très grand nombre si le modèle comporte plusieurs équations et, pour des raisons expliquées dans la prochaine section, on peut ne pas vouloir utiliser un nombre aussi grand d'instruments. Donc, en pratique, il existe plusieurs moyens valables de choisir  $\mathbf{W}$ .

Il existe deux objectifs opposés dans le choix de  $\mathbf{W}$ . D'une part, nous aimerions obtenir des estimations qui soient aussi efficaces que possible asymptotiquement. D'autre part, nous aimerions obtenir des estimations qui aient le plus petit biais possible pour un échantillon fini. Malheureusement ces deux objectifs s'avèrent incompatibles l'un avec l'autre. Nous discuterons ici de ce problème d'efficacité asymptotique et du problème des propriétés avec un échantillon fini dans la prochaine section.

Supposons qu'il y ait deux choix possibles de matrice d'instruments,  $\mathbf{W}_1$  et  $\mathbf{W}_2$ , où  $\mathbf{W}_2$  est composé de  $\mathbf{W}_1$  et d'au moins une autre colonne, ce qui implique que  $\mathcal{S}(\mathbf{W}_1)$  est un sous-espace de  $\mathcal{S}(\mathbf{W}_2)$ . Les estimateurs IV qui en résultent sont

$$\tilde{\beta}^1 = (\mathbf{X}^\top \mathbf{P}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_1 \mathbf{y} \quad \text{et}$$

$$\tilde{\beta}^2 = (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_2 \mathbf{y},$$

où  $\mathbf{P}_1$  et  $\mathbf{P}_2$  désignent les matrices qui projettent orthogonalement sur les espaces engendrés par  $\mathbf{W}_1$  et  $\mathbf{W}_2$ . Les matrices de covariance asymptotiques de ces deux estimateurs sont les suivantes:

$$\mathbf{V}^\infty(n^{1/2}(\tilde{\beta}^i - \beta_0)) = \sigma_0^2 \operatorname{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{P}_i \mathbf{X})^{-1} \quad (7.26)$$

pour  $i = 1, 2$ . Nous pouvons donc conclure que  $\tilde{\beta}^2$  est au moins aussi efficace que  $\tilde{\beta}^1$  si la différence entre leurs matrices de covariance asymptotiques est semi-définie positive. Cela est bien le cas, comme nous allons le démontrer.

Considérons la différence

$$\mathbf{X}^\top \mathbf{P}_2 \mathbf{X} - \mathbf{X}^\top \mathbf{P}_1 \mathbf{X} = \mathbf{X}^\top (\mathbf{P}_2 - \mathbf{P}_1) \mathbf{X}. \quad (7.27)$$

Puisque  $\mathcal{S}(\mathbf{W}_1)$  est un sous-espace de  $\mathcal{S}(\mathbf{W}_2)$ ,  $\mathbf{P}_1 \mathbf{P}_2 = \mathbf{P}_2 \mathbf{P}_1 = \mathbf{P}_1$ , et par conséquent  $\mathbf{P}_2 - \mathbf{P}_1$  est une projection orthogonale. Il en résulte que (7.27) est semi-définie positive. Dans l'Annexe A, nous montrons que la différence entre deux matrices symétriques, définies positive est semi-définie positive si et seulement si la différence des opposées de leurs inverses est semi-définie positive. Par conséquent,

$$(\mathbf{X}^\top \mathbf{P}_1 \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1}$$

est aussi semi-définie positive. D'où, à partir de (7.26), la différence entre les matrices de covariance asymptotiques de  $\tilde{\beta}^1$  et  $\tilde{\beta}^2$  est semi-définie positive. Nous pouvons donc conclure que  $\tilde{\beta}^2$  est asymptotiquement aussi efficace que  $\tilde{\beta}^1$ . Cela est cohérent puisque  $\mathbf{W}_2$  explique  $\mathbf{X}$  au moins aussi bien que  $\mathbf{W}_1$ .

Il existe un cas particulier pour lequel  $\tilde{\beta}^2$  et  $\tilde{\beta}^1$  sont d'une efficacité asymptotique identique et tendent en effet vers le même vecteur aléatoire lorsque  $n \rightarrow \infty$ . Ce cas survient lorsque  $\mathbf{W}_2$  a le même pouvoir explicatif, asymptotiquement que  $\mathbf{W}_1$  sur  $\mathbf{X}$ . Cela se produira si, par exemple,  $\mathbf{W}_1$  se compose de toutes les variables exogènes et prédéterminées d'un modèle linéaire à équations simultanées, parce que les variables supplémentaires dans  $\mathbf{W}_2$  ne doivent pas avoir de pouvoir explicatif supplémentaire pour  $\mathbf{X}$ . Mais cela est un cas très particulier. Dans chaque autre cas,  $\tilde{\beta}^2$  est asymptotiquement plus efficace que  $\tilde{\beta}^1$ .

Ce résultat semble nous suggérer que nous devrions utiliser autant d'instruments que possible. Cela est vrai si  $n$  est très grand, et quand les propriétés asymptotiques sont les seules choses qui nous intéressent. Mais ce n'est pas une bonne démarche à suivre pour des échantillons de tailles modérées. Le problème est que l'accroissement du nombre des instruments tend à rendre le biais des estimateurs IV, avec des échantillons finis, de plus en plus important, et c'est de ce sujet dont nous allons parler dans la section qui suit.

## 7.5 LES DOUBLES MOINDRES CARRÉS

Ce à quoi nous nous sommes référés en tant qu'estimateur IV de  $\beta$  dans le modèle de régression linéaire (7.01),  $\tilde{\beta}$ , est aussi connu sous le nom d'estimateur des **doubles moindres carrés**, ou **2SLS**. A l'origine, il fut proposé par Theil (1953) et indépendamment par Basman (1957), dans le contexte du modèle à équations simultanées. Le nom "doubles moindres carrés" évoque une méthode particulière par laquelle cet estimateur IV particulier peut être calculé, et cette terminologie est donc généralement utilisée en économétrie. Cependant, l'idée principale qui sous-tend l'estimation IV est beaucoup plus générale que celle de l'estimation 2SLS. Comme nous le verrons dans la prochaine section, la procédure IV se généralise par exemple naturellement au cas des modèles de régression non linéaire, ce qui n'est pas le cas des 2SLS. Par conséquent, nous préférons insister sur l'interprétation de l'estimateur  $\tilde{\beta}$  par IV plutôt que par 2SLS.

Les doubles moindres carrés s'opèrent de la manière suivante. Dans la première étape, toutes les variables explicatives endogènes courantes d'un système d'équations simultanées sont régressées sur la matrice d'instruments  $\mathbf{W}$ . Dans la seconde étape, chaque équation est estimée par OLS après que toutes les variables endogènes qui apparaissent du côté droit aient été remplacées par les valeurs ajustées des régressions correspondantes de la première étape. Par conséquent, pour chaque équation structurelle du système, la variable endogène située à gauche de l'équation, est régressée sur un ensemble

de régresseurs qui contient les variables exogènes et prédéterminées qui apparaissent du côté droit de l'équation, plus les valeurs ajustées des variables explicatives endogènes dans cette équation provenant des régressions de la première étape.

Si  $\mathbf{y}$  désigne une des variables endogènes du système, si  $\mathbf{X}$  désigne l'ensemble des variables explicatives, endogènes et exogènes, qui sont associées à  $\mathbf{y}$  dans l'équation, et si  $\mathbf{W}$  désigne l'ensemble de toutes les variables exogènes et prédéterminées du système global, la régression de  $\mathbf{y}$  de la seconde étape peut s'écrire de façon plus simple comme

$$\mathbf{y} = \mathbf{P}_W \mathbf{X} \boldsymbol{\beta} + \text{résidus.} \quad (7.28)$$

L'estimateur OLS de  $\boldsymbol{\beta}$  de cette régression n'est autre que l'estimateur IV (7.17):

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y}.$$

Cependant, nous remarquons que l'estimation OLS de la matrice de covariance (7.28) n'est pas celle que nous voulons. Cette estimation sera

$$\frac{\|\mathbf{y} - \mathbf{P}_W \mathbf{X} \tilde{\boldsymbol{\beta}}\|^2}{n - k} (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1}, \quad (7.29)$$

tandis que l'estimation (7.24) qui a été déduite auparavant peut être réécrite comme

$$\frac{\|\mathbf{y} - \mathbf{X} \tilde{\boldsymbol{\beta}}\|^2}{n} (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1}. \quad (7.30)$$

Ces deux estimations ne sont pas les mêmes. Elles seraient les mêmes seulement si les procédures OLS et IV étaient identiques, c'est-à-dire si  $\mathbf{X} = \mathbf{P}_W \mathbf{X}$ . De plus,  $n$  devrait être remplacé par  $n - k$  dans (7.30). Le problème est que la régression par OLS de la seconde étape ne nous fournit pas une bonne estimation de  $\sigma^2$ ; elle utilise  $\mathbf{y} - \mathbf{P}_W \mathbf{X} \tilde{\boldsymbol{\beta}}$  plutôt que  $\mathbf{y} - \mathbf{X} \tilde{\boldsymbol{\beta}}$  comme vecteur de résidus. Les résidus de la seconde étape  $\mathbf{y} - \mathbf{P}_W \mathbf{X} \tilde{\boldsymbol{\beta}}$  tendront à être trop importants asymptotiquement puisque  $\mathbf{P}_W \mathbf{X}$  aura moins de pouvoir explicatif que  $\mathbf{X}$  elle-même si le modèle est correctement spécifié. Cela bien sûr, peut ne pas être vrai avec des échantillons finis, ou si le modèle est correctement spécifié. Si nous exécutons effectivement les 2SLS en deux étapes, plutôt que de compter sur une procédure 2SLS ou IV pré-programmée, nous devons faire attention à utiliser (7.30) plutôt que (7.29) pour la matrice de variance estimée.<sup>2</sup> Des programmes pour l'estimation 2SLS remplacent normalement

<sup>2</sup> Les 2SLS sont un cas spécial d'une régression de ce que Pagan (1984b, 1986) appelle "régresseurs générés". Même quand les régressions offrent des paramètres estimés convergents, ils donnent habituellement des estimations non convergentes de la matrice de covariance des paramètres estimés. La non convergence de (7.29) nous donne un simple exemple de ce phénomène.

$P_W \mathbf{X} \tilde{\beta}$  par  $\mathbf{X} \tilde{\beta}$  avant de calculer la somme des carrés expliquée, la somme des carrés des résidus, le  $R^2$ , et d'autres statistiques qui dépendent de ces quantités.

Il y a eu une grande quantité de travaux sur les propriétés des 2SLS avec des échantillons finis, c'est-à-dire l'estimateur IV  $\tilde{\beta}$ . Parmi ces travaux, nous pouvons citer Anderson (1982), Anderson et Sawa (1979), Mariano (1982), Phillips (1983), et Taylor (1983). Malheureusement, beaucoup de résultats issus de cette littérature ne sont spécifiques qu'aux modèles développés. Un des résultats importants attribué à Kinal, est que le  $m^{\text{ième}}$  moment de l'estimateur 2SLS existe si et seulement si

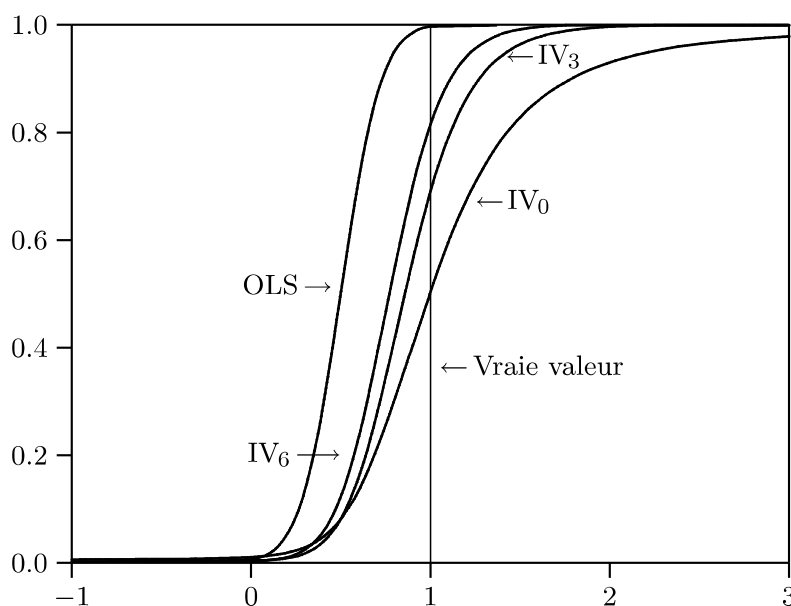
$$m < l - k + 1.$$

Le membre de droite représente ici la différence entre le nombre d'instruments et le nombre de régresseurs, plus un. C'est aussi le degré de suridentification de l'équation plus un; voir la Section 7.8. D'après ce résultat, l'estimateur 2SLS n'aura même pas d'espérance si  $l = k$  (dans le cas où l'estimateur 2SLS se réduit à l'estimateur IV simple discuté dans la section précédente). Cela laisse supposer que ses propriétés avec des échantillons finis seront plus pauvres dans ce cas, et en effet, elles le sont souvent; voir Nelson et Startz (1990a, 1990b). Puisque, en général, nous voulons des estimateurs avec au moins une espérance et une variance, le bon sens voudrait que, si possible,  $l$  soit toujours supérieur à  $k$  d'au moins 2.

En fait, le résultat de l'efficacité asymptotique de la section précédente suggère que  $l$  doit être aussi grand que possible. Cependant, la théorie avec des échantillons finis et les expériences Monte Carlo suggèrent que ce n'est pas toujours une bonne idée. Le problème fondamental est que lorsque nous ajoutons de plus en plus de colonnes à  $\mathbf{W}$ , cette dernière explique de mieux en mieux les colonnes de  $\mathbf{X}$  qui ne se situent pas dans  $\mathcal{S}(\mathbf{W})$ , et  $P_W \mathbf{X}$  ressemble de plus en plus à  $\mathbf{X}$ . C'est une conséquence inévitable de la tendance des OLS d'ajuster trop bien. Par conséquent, l'estimateur IV tend à devenir de plus en plus biaisé quand  $l$  augmente, et il approche en fin de compte l'estimateur OLS lorsque  $l$  approche  $n$ .

La Figure 7.1 nous en donne une illustration. Elle montre les fonctions de répartition de l'estimateur OLS et de trois estimateurs IV différents dans un cas simple. Les trois estimateurs IV, notés  $IV_0$ ,  $IV_3$ , et  $IV_6$ , ont respectivement  $l - k$  égal à 0, 3, et 6. La quantité qui est estimée est le paramètre de la pente d'une équation avec un régresseur endogène et un terme constant; sa vraie valeur est l'unité. La taille de l'échantillon est seulement de 25 de façon à faire apparaître les biais pour un échantillon fini. Ces fonctions de répartition ont été estimées au moyen d'une expérience Monte Carlo (voir le Chapitre 21), les détails étant peu importants puisque la figure n'est que purement illustrative.

La courbe la plus à gauche dans la Figure 7.1 illustre la fonction de répartition de l'estimateur OLS qui, dans ce cas, est sévèrement biaisé vers le

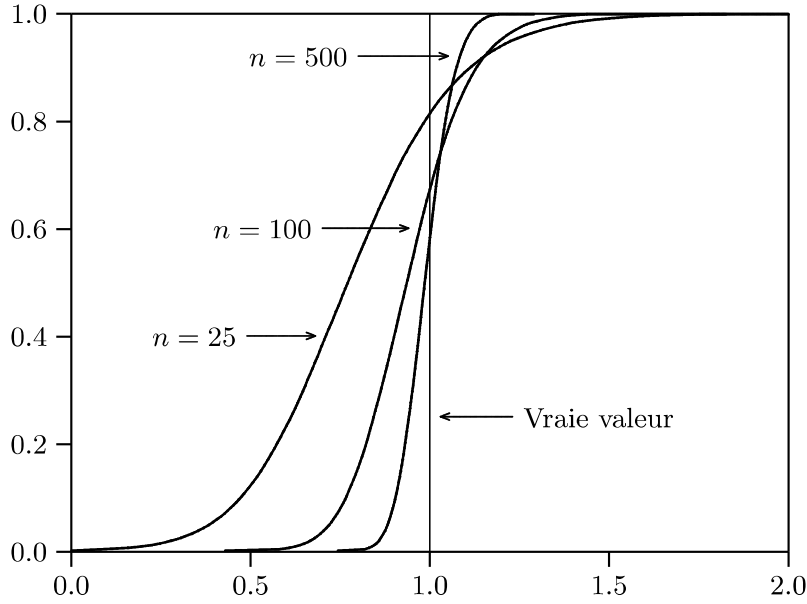


**Figure 7.1** Distributions des estimations OLS et IV,  $n = 25$

bas. La courbe la plus à droite est celle de  $IV_0$  qui a approximativement la vraie médiane, mais qui a aussi beaucoup plus de dispersion (nous ne pouvons pas dire variance puisqu'il n'y a pas de second moment) et des queues plus épaisses que l'estimateur OLS. En effet, parmi les 50,000 simulations que nous avons effectuées, nous avons obtenu plusieurs estimations  $IV_0$  plus grandes que 1000 en valeur absolue! Les fonctions de répartition pour  $IV_3$  et  $IV_6$  sont pour l'essentiel comprises entre celles de OLS et de  $IV_0$ , et ont des queues beaucoup plus fines que ces dernières, avec  $IV_6$  beaucoup plus proche de OLS que  $IV_3$ , comme la discussion antérieure le prédisait. Ces deux estimateurs sont assez sévèrement biaisés (se rappeler qu'ici  $n = 25$ ), bien qu'ils le soient moins que celui des OLS, mais ont évidemment plus de variance que OLS compte tenu des pentes moins raides de leurs fonctions de répartition.

Le meilleur estimateur dépend du critère qui est utilisé pour choisir parmi tous les estimateurs. Si on ne s'intéresse qu'à la médiane,  $IV_0$  est clairement le meilleur. Mais si d'un autre côté nous utilisons l'erreur quadratique moyenne comme critère,  $IV_0$  est clairement le plus mauvais, puisque n'ayant pas de premier moment ou de moments supérieurs, son erreur quadratique moyenne est infinie. Fondé sur la plupart des critères, le choix se porterait sur  $IV_3$  ou  $IV_6$ . Pour un  $n$  assez grand, le second serait bien sûr préférable, puisque son plus grand biais disparaîtra quand  $n$  augmentera, alors que sa plus petite variance persistera. L'effet d'un échantillon de plus en plus grand est illustré dans la Figure 7.2 qui montre la répartition de  $IV_6$  pour  $n = 25$ ,  $n = 100$ , et  $n = 500$ . Etant donné que  $n$  augmente, à la fois la variance et le biais de l'estimateur diminuent, comme prévu, même si pour  $n = 500$ , le biais est non négligeable.





**Figure 7.2** Distributions des estimations  $IV_6$  pour plusieurs tailles d'échantillon

Il faudrait souligner que les Figures 7.1 et 7.2 s'appliquent seulement à un exemple particulier, dans lequel les instruments s'avèrent être généralement de bons instruments. Dans d'autres cas, et spécialement quand les instruments possèdent une faible capacité à expliquer les régresseurs endogènes, les estimateurs IV peuvent être extrêmement inefficaces, et leurs distributions avec un échantillon fini peuvent être très différentes de leurs distributions asymptotiques.

## 7.6 LES VARIABLES INSTRUMENTALES: LE CAS NON LINÉAIRE

Il est facile de généraliser la procédure IV linéaire discutée antérieurement au cas des modèles de régression non linéaire. Supposons que le modèle soit

$$y_t = x_t(\beta) + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad (7.31)$$

où la fonction de régression  $x_t(\beta)$  dépend implicitement des variables endogènes communes aussi bien que des variables exogènes et prédéterminées, et  $\beta$  est un vecteur à  $k$  composantes, comme d'habitude. Supposons que l'on dispose d'une matrice d'instruments valables  $\mathbf{W}$ , avec  $l \geq k$  comme auparavant, l'objectif étant de minimiser seulement la partie de la distance entre  $\mathbf{y}$  et  $\mathbf{x}(\beta)$  qui se trouve dans  $\mathcal{S}(\mathbf{W})$ . Cela peut être réalisé en minimisant la fonction critère

$$\|P_W(\mathbf{y} - \mathbf{x}(\beta))\|^2 = (\mathbf{y} - \mathbf{x}(\beta))^\top P_W(\mathbf{y} - \mathbf{x}(\beta)). \quad (7.32)$$

Cette fonction critère est l'équivalent non linéaire de (7.15). Les conditions du premier ordre qui caractérisent les estimations IV  $\tilde{\beta}$  sont

$$\mathbf{X}^\top(\tilde{\beta})\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\tilde{\beta})) = \mathbf{0}, \quad (7.33)$$

où, comme d'habitude, la matrice  $\mathbf{X}(\beta)$  de dimension  $n \times k$  a pour élément type

$$X_{ti}(\beta) = \frac{\partial x_t(\beta)}{\partial \beta_i}.$$

Les conditions (7.33) sont à l'évidence les conditions analogues, dans le cas non linéaire, aux conditions du premier ordre (7.16) dans le cas linéaire. Elles indiquent que les résidus  $\mathbf{y} - \mathbf{x}(\tilde{\beta})$  doivent être orthogonaux à la matrice des dérivées  $\mathbf{X}(\tilde{\beta})$ , après que cette dernière eût été projetée sur  $\mathcal{S}(\mathbf{W})$ . Comme pour le cas NLS, nous ne pouvons espérer résoudre analytiquement les équations de (7.33) pour trouver  $\tilde{\beta}$ , bien que cela puisse être possible dans des cas particuliers.

La démonstration de la convergence et de la normalité asymptotique des estimations IV non linéaires  $\tilde{\beta}$  est assez directe, compte tenu de certaines conditions de régularités adéquates. Parmi celles-ci, les principales dont nous avons besoin sont celles qui permettent aux NLS d'être convergents et asymptotiquement normaux (voir les Sections 5.3 et 5.4), mais aussi des versions modifiées des hypothèses (7.18a), (7.18b) et (7.18c) en remplacement de l'hypothèse d'indépendance entre les aléas et la fonction de régression et ses dérivées. Dans la dernière de ces hypothèses, la matrice  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$  remplace la matrice  $\mathbf{X}$ , où  $\beta_0$  est, comme d'habitude, la valeur de  $\beta$  correspondant au DGP, qui est supposé être un cas particulier du modèle que nous estimons. Le dernier résultat est

$$n^{1/2}(\tilde{\beta} - \beta_0) \overset{a}{\sim} N\left(\mathbf{0}, \sigma_0^2 \text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}_0^\top \mathbf{P}_W \mathbf{X}_0)^{-1}\right), \quad (7.34)$$

qui ressemble étroitement à (7.23) dans le cas linéaire.

L'**estimateur IV non linéaire** basé sur la minimisation de la fonction critère (7.32) fut proposé par Amemiya (1974), qui de manière trompeuse, l'appela **l'estimateur des doubles moindres carrés non linéaires**, ou **NL2SLS**. En fait, il ne se calcule *pas du tout* en deux étapes. En essayant de calculer un estimateur analogue à celui des 2SLS linéaires, cela donnerait, en général, un estimateur non convergent très différent de celui des IV non linéaires.

Il est intéressant de voir pourquoi il en est ainsi. Nous devons rendre explicite la dépendance de  $\mathbf{x}(\beta)$  aux variables explicatives. Par conséquent, le modèle (7.31) peut se récrire comme

$$\mathbf{y} = \mathbf{x}(\mathbf{Z}, \beta) + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

où  $\mathbf{x}(\mathbf{Z}, \beta)$  est un vecteur avec un élément type  $x_t(\mathbf{Z}_t, \beta)$ ,  $\mathbf{Z}$  étant une matrice des observations sur les variables explicatives, dont la ligne  $t$  est  $\mathbf{Z}_t$  et dont

certaines colonnes peuvent être corrélées avec  $\mathbf{u}$ . La matrice  $\mathbf{Z}$  n'est pas nécessairement de dimension  $n \times k$ , parce qu'il peut y avoir plus ou moins de paramètres que de variables explicatives. Une procédure 2SLS régresserait ces colonnes de  $\mathbf{Z}$ , qui sont potentiellement corrélées avec  $\mathbf{u}$ , sur la matrice des instruments  $\mathbf{W}$  de façon à obtenir  $\mathbf{P}_W \mathbf{Z}$ . Elle minimiserait alors la fonction objectif

$$(\mathbf{y} - \mathbf{x}(\mathbf{P}_W \mathbf{Z}, \boldsymbol{\beta}))^\top (\mathbf{y} - \mathbf{x}(\mathbf{P}_W \mathbf{Z}, \boldsymbol{\beta})). \quad (7.35)$$

Cette procédure donnerait des estimations convergentes si les fonctions de régression  $x_t(\mathbf{Z}_t, \boldsymbol{\beta})$  étaient linéaires en tous les éléments endogènes de  $\mathbf{Z}_t$ . Mais si les fonctions de régression étaient non linéaires en au moins un des éléments de  $\mathbf{Z}_t$ , minimiser (7.35) ne donnerait pas des estimations convergentes, parce que même si  $\mathbf{P}_W \mathbf{Z}$  était asymptotiquement orthogonale à  $\mathbf{u}$ ,  $\mathbf{x}(\mathbf{P}_W \mathbf{Z}, \boldsymbol{\beta})$  ne le serait pas.

Pour considérer un exemple très simple, supposons que la fonction de régression  $x_t(\mathbf{Z}_t, \boldsymbol{\beta})$  soit  $\beta z_t^2$ . Par conséquent, ce serait juste une variable indépendante qui est corrélée avec  $u_t$ , et un paramètre. La théorie pour les régressions linéaires est applicable pour cet exemple, puisque la fonction de régression est linéaire en son paramètre  $\beta$ . L'obtention d'une estimation convergente de  $\beta$  passe par la minimisation de  $\|\mathbf{P}_W(\mathbf{y} - \beta \mathbf{z}^2)\|^2$  par rapport à  $\beta$ , où  $\mathbf{z}^2$  désigne le vecteur dont l'élément type est  $z_t^2$ . A l'opposé, si on projetait en premier  $\mathbf{z}$  sur  $\mathbf{W}$  à l'aide d'une procédure 2SLS, nous minimiserions  $\|\mathbf{y} - \beta(\mathbf{P}_W \mathbf{z})^2\|^2$ , où  $(\mathbf{P}_W \mathbf{z})^2$  serait le vecteur avec l'élément type  $(\mathbf{P}_W \mathbf{z})_t^2$ . La deuxième minimisation n'est évidemment pas limitée au sous-espace  $\mathcal{S}(\mathbf{W})$  et donc, ne conduira pas, en général, à des estimations convergentes de  $\beta$ .

Dans de nombreux cas, le plus grand problème avec les procédures IV non linéaires concerne le choix de  $\mathbf{W}$ . Avec un modèle linéaire, il est relativement facile de choisir  $\mathbf{W}$ . Si l'équation à estimer provient d'un système d'équations linéaires simultanées, nous savons d'après la forme réduite que toutes les variables endogènes dépendent de façon linéaire des variables exogènes et prédéterminées. Par conséquent, une chose naturelle à effectuer est de prendre  $\mathbf{W}$  composée de toutes les variables exogènes et prédéterminées du modèle, sauf s'il y en a beaucoup trop. Quand un modèle est non linéaire en ses variables endogènes, cette approche ne peut plus fonctionner. Il n'existe pas de forme réduite non contrainte comparable à celle des modèles linéaires. Les variables endogènes peuvent dépendre des variables exogènes et prédéterminées en des formes non linéaires. Cela suggère d'utiliser les puissances et les produits croisés des variables exogènes comme instruments. Mais il n'est pas évident de savoir combien de puissances ou de produits croisés il faut utiliser, et le problème d'avoir beaucoup trop d'instruments est difficile à traiter, même quand le nombre de variables exogènes et prédéterminées est petit. Pour en savoir plus, voir Amemiya (1974), Kelejian (1971), et Bowden et Turkington (1984, Chapitre 5).

## 7.7 LES TESTS D'HYPOTHÈSES BASÉS SUR LA GNR

Comme nous l'avons vu dans le Chapitre 6, nous associons une **régression de Gauss-Newton** à tout modèle de régression non linéaire estimé par moindres carrés. Il en est de même pour chaque modèle de régression estimé par variables instrumentales. Pour ces derniers la forme générale de la GNR est

$$\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}^*) = \mathbf{P}_W \mathbf{X}(\boldsymbol{\beta}^*) \mathbf{b} + \text{résidus}, \quad (7.36)$$

où  $\boldsymbol{\beta}^*$  peut être n'importe quelle valeur spécifiée de  $\boldsymbol{\beta}$ . Par conséquent, la seule différence entre cette GNR et la GNR originale, est que les régresseurs sont multipliés par  $\mathbf{P}_W$ . Cette variante de la GNR possède presque toutes les mêmes propriétés que la GNR d'origine étudiée au Chapitre 6. Comme cette dernière, elle peut être utilisée pour une variété d'usages en fonction de son point d'évaluation.

Si la GNR (7.36) est évaluée en les estimations IV  $\tilde{\boldsymbol{\beta}}$ , l'estimation OLS  $\tilde{\mathbf{b}}$  sera identiquement égale à zéro. Comme d'habitude, cela peut fournir une voie intéressante pour vérifier la précision du programme d'optimisation non linéaire employé. De plus, l'estimation OLS de la matrice de covariance à partir de la régression artificielle nous fournira une estimation valable de la matrice de covariance de  $\tilde{\boldsymbol{\beta}}$ . Parce que  $\mathbf{P}_W \mathbf{X}(\tilde{\boldsymbol{\beta}})$  ne peut pas avoir de pouvoir explicatif sur  $\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})$ , la somme des carrés des résidus doit être simplement  $\|\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})\|^2$ , et l'estimation OLS de la matrice de covariance sera donc

$$\frac{\|\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})\|^2}{n - k} (\mathbf{X}^\top(\tilde{\boldsymbol{\beta}}) \mathbf{P}_W \mathbf{X}(\tilde{\boldsymbol{\beta}}))^{-1}. \quad (7.37)$$

Cette expression est l'analogue non linéaire de l'expression (7.30), excepté que nous trouvons maintenant  $n - k$  au dénominateur de l'estimation de  $\sigma^2$  au lieu de  $n$ . Elle offre, à l'évidence, une manière valable d'estimer l'analogue avec des échantillons finis de la matrice de covariance asymptotique qui apparaît dans (7.34). Il existe malgré tout un doute sur la pertinence de l'ajustement des degrés de liberté, ainsi que nous l'avons remarqué avant (7.24), mais l'expression (7.37) est tout de même exactement ce que nous voulons utiliser pour estimer la matrice de covariance de  $\tilde{\boldsymbol{\beta}}$ .

La GNR (7.36) peut être utilisée, bien sûr, pour d'autres usages. Si elle est évaluée en des estimations convergentes mais non efficaces, elle peut être utilisée pour calculer des estimations efficaces en une étape, qui sont asymptotiquement équivalentes à  $\tilde{\boldsymbol{\beta}}$ ; voir la Section 6.6. Elle peut être aussi utilisée comme partie d'une procédure d'optimisation numérique pour minimiser la fonction critère (7.32); voir la Section 6.8. Dans les deux cas, il n'existe pas de différence majeure entre les résultats pour les versions NLS et IV de la GNR. Cependant, l'application principale des régressions de Gauss-Newton consiste probablement à calculer des statistiques de test basées sur les principes du multiplicateur de Lagrange et  $C(\alpha)$ , c'est-à-dire tester des contraintes sur  $\boldsymbol{\beta}$

sans avoir recours à une estimation non contrainte du modèle. Puisque le cas IV est un peu différent du cas NLS, ce sujet mérite une discussion.

Supposons que  $\check{\beta}$  soit un vecteur d'estimations IV sujet à un ensemble de  $r$  restrictions éventuellement non linéaires. Pour simplifier l'exposé, nous supposons que le modèle est paramétrisé pour que  $\mathbf{x}(\beta) \equiv \mathbf{x}(\beta_1, \beta_2)$ , où  $\beta_1$  est de dimension  $(k-r) \times 1$  et  $\beta_2$  est de dimension  $r \times 1$ , et que les restrictions soient  $\beta_2 = \mathbf{0}$ . Cependant, puisque la manière d'écrire ces contraintes n'est qu'une question de paramétrisation, les résultats seraient les mêmes si nous tenions compte des contraintes non linéaires générales de la forme  $\mathbf{r}(\beta) = \mathbf{0}$ .

Quand nous évaluons la GNR (7.36) avec les estimations contraintes  $\check{\beta}$ , elle devient

$$\mathbf{y} - \check{\mathbf{x}} = \mathbf{P}_W \check{\mathbf{X}} \mathbf{b} + \text{résidus}, \quad (7.38)$$

où  $\check{\mathbf{x}} \equiv \mathbf{x}(\check{\beta})$  et  $\check{\mathbf{X}} \equiv \mathbf{X}(\check{\beta})$ . Cette régression artificielle génère des statistiques de test de la même manière que la GNR dans le cas des moindres carrés non linéaires. La somme des carrés expliquée de (7.38) est

$$(\mathbf{y} - \check{\mathbf{x}})^\top \mathbf{P}_W \check{\mathbf{X}} (\check{\mathbf{X}}^\top \mathbf{P}_W \check{\mathbf{X}})^{-1} \check{\mathbf{X}}^\top \mathbf{P}_W (\mathbf{y} - \check{\mathbf{x}}). \quad (7.39)$$

Quand la somme des carrés expliquée est divisée par n'importe quelle estimation convergente de  $\sigma^2$ , le résultat est asymptotiquement distribué, sous l'hypothèse nulle, suivant une  $\chi^2(r)$ . Une statistique de test valable peut être obtenue en calculant  $n$  fois le  $R^2$  non centré de la GNR (7.38). D'autres statistiques de test seront abordées ultérieurement. Nous remarquons que le  $R^2$  est celui des OLS, et non pas celui d'une procédure IV qui serait utilisée si l'on régressait  $\mathbf{y} - \check{\mathbf{x}}$  sur  $\check{\mathbf{X}}$  en utilisant  $\mathbf{W}$  comme matrice d'instruments.

Afin de comprendre pourquoi les statistiques de test basées sur la GNR sont valables, il est pratique de récrire (7.38) en partitionnant le vecteur de paramètres  $\beta$  en  $\beta_1$  et  $\beta_2$ :

$$\mathbf{y} - \check{\mathbf{x}} = \mathbf{P}_W \check{\mathbf{X}}_1 \mathbf{b}_1 + \mathbf{P}_W \check{\mathbf{X}}_2 \mathbf{b}_2 + \text{résidus}, \quad (7.40)$$

où  $\check{\mathbf{X}}_i$  contient les colonnes de  $\check{\mathbf{X}}$  qui correspondent à  $\beta_i$  pour  $i = 1, 2$ . Puisque  $\check{\mathbf{X}}_1$  ne peut pas avoir de pouvoir explicatif sur  $\mathbf{y} - \check{\mathbf{x}}$  d'après les conditions du premier ordre pour les estimations contraintes, le Théorème FWL implique que l'équation (7.40) doit avoir la même somme des carrés expliquée que

$$\check{\mathbf{M}}_1 (\mathbf{y} - \check{\mathbf{x}}) = \check{\mathbf{M}}_1 \mathbf{P}_W \check{\mathbf{X}}_2 \mathbf{b}_2 + \text{résidus},$$

où

$$\check{\mathbf{M}}_1 \equiv \mathbf{I} - \mathbf{P}_W \check{\mathbf{X}}_1 (\check{\mathbf{X}}_1^\top \mathbf{P}_W \check{\mathbf{X}}_1)^{-1} \check{\mathbf{X}}_1^\top \mathbf{P}_W$$

est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{P}_W \check{\mathbf{X}}_1)$ . Par conséquent, nous voyons que la somme des carrés expliquée (7.39) peut se récrire comme

$$(\mathbf{y} - \check{\mathbf{x}})^\top \check{\mathbf{M}}_1 \mathbf{P}_W \check{\mathbf{X}}_2 (\check{\mathbf{X}}_2^\top \mathbf{P}_W \check{\mathbf{M}}_1 \mathbf{P}_W \check{\mathbf{X}}_2)^{-1} \check{\mathbf{X}}_2^\top \mathbf{P}_W \check{\mathbf{M}}_1 (\mathbf{y} - \check{\mathbf{x}}). \quad (7.41)$$

Nous ne démontrerons pas que l'expression (7.41), quand elle est divisée par une estimation convergente de  $\sigma^2$ , est asymptotiquement distribuée suivant une  $\chi^2(r)$ . La preuve résulte étroitement de celle employée dans le cas des moindres carrés non linéaires qui a été esquissée dans la Section 6.4. Évidemment, le résultat s'ensuit immédiatement si nous pouvons montrer que le vecteur à  $r$  composantes

$$n^{-1/2} \tilde{\mathbf{X}}_2^\top \mathbf{P}_W \tilde{\mathbf{M}}_1 (\mathbf{y} - \tilde{\mathbf{x}})$$

suit asymptotiquement une loi normale avec une matrice de covariance

$$\sigma_0^2 \operatorname{plim}_{n \rightarrow \infty} (n^{-1} \tilde{\mathbf{X}}_2^\top \mathbf{P}_W \tilde{\mathbf{M}}_1 \mathbf{P}_W \tilde{\mathbf{X}}_2).$$

Il peut être un bon exercice pour les lecteurs de démontrer ce résultat. Pour une discussion plus détaillée, voir Engle (1982a).

Les tests qui viennent d'être décrits sont basés sur le principe LM. Il est, bien sûr, aussi valable d'utiliser les tests basés sur le principe  $C(\alpha)$ , discuté dans la Section 6.7. Les régressions de test ressembleraient à (7.40), excepté que la régressande et les régresseurs seraient évalués en des estimations  $\hat{\beta}$  qui sont convergentes sous l'hypothèse nulle mais pour lesquelles les conditions du premier ordre ne sont pas satisfaites:

$$(\mathbf{y} - \hat{\mathbf{x}})^\top \mathbf{P}_W \hat{\mathbf{X}}_1 \neq \mathbf{0}.$$

Par conséquent, la régressande de la GNR ne serait plus orthogonale aux régresseurs qui correspondent à  $\beta_1$ , et les tests basés sur la somme des carrés expliquée ne seraient pas valables. Mais n'importe quel test basé sur la réduction de la SSR provoquée par ajout des régresseurs qui correspondent à  $\beta_2$ , tel qu'un test pseudo- $F$ , serait sûrement valable. L'approche la plus facile est simplement d'exécuter deux fois la GNR, une fois avec et une fois sans les régresseurs qui correspondent à  $\beta_2$ , et calculer alors un test de Fisher comme d'habitude.

Ce type de test peut être particulièrement utile quand la matrice des instruments qui devrait être utilisée pour estimer l'hypothèse alternative possède plus de colonnes que celle effectivement utilisée pour estimer l'hypothèse nulle. Cela peut facilement se produire si la fonction de régression pour le modèle alternatif dépend d'une ou de plusieurs variables exogènes et prédéterminées qui n'apparaissent pas dans la matrice d'instruments du modèle correspondant à l'hypothèse nulle. Pour calculer un test basé sur le principe LM, nous devrions revenir et estimer encore le modèle correspondant à l'hypothèse nulle en utilisant la même matrice d'instruments utilisée pour estimer le modèle correspondant à l'hypothèse alternative. Cette étape n'est pas nécessaire si nous utilisons un test basé sur le principe  $C(\alpha)$ . Nous utilisons malgré tout la matrice d'instruments la plus importante pour construire

les régresseurs de la GNR, et par conséquent, la régressande ne sera plus orthogonale aux régresseurs qui correspondent à  $\beta_1$ , mais cela n'affecte pas la validité de la statistique de test.

Toute la discussion précédente suppose que la GNR (7.38) est calculée par OLS. Cependant en pratique, il semblerait plus facile de régresser  $\mathbf{y} - \tilde{\mathbf{x}}$  sur  $\tilde{\mathbf{X}}$  par une procédure IV en utilisant  $\mathbf{W}$  comme matrice d'instruments. Bien que cela évite l'étape initiale de régresser les colonnes de  $\tilde{\mathbf{X}}$  sur  $\mathbf{W}$ , ce n'est pas une bonne idée. Nous ne pourrions pas utiliser la somme des carrés expliquée reportée par le progiciel pour calculer les statistiques de test avec plus d'un degré de liberté, puisque ce ne serait pas vraiment la somme des carrés expliquée de la régression (7.38) (voir la Section 7.5). Pour la même raison, nous ne pouvons pas construire des tests pseudo- $F$  en utilisant les sommes des carrés des résidus obtenues à partir d'une estimation IV d'un modèle contraint et d'un modèle non contraint.

Nous avons vu que si la régression de Gauss-Newton est exécutée par les OLS, nous pouvons tester des contraintes sur  $\beta$  précisément de la même manière que dans le contexte des moindres carrés non linéaires. Cependant, utiliser d'autres méthodes peut être quelque peu différent. La raison est que  $\sigma^2$  doit être estimée, et comme nous l'avons vu dans la Section 7.5, il peut être délicat d'estimer  $\sigma^2$  en utilisant les IV. Dans le cas de la GNR, il est clair que l'on peut estimer  $\sigma^2$  de façon valide par

$$\frac{1}{n}(\mathbf{y} - \tilde{\mathbf{x}})^\top(\mathbf{y} - \tilde{\mathbf{x}}),$$

où  $n$  devrait être remplacé par  $(n - k + r)$ . Cette estimation de  $\sigma^2$  est basée sur les estimations contraintes. Il est aussi valable d'utiliser  $n^{-1}$  ou  $(n - k)^{-1}$  fois la somme des carrés des résidus de la GNR (7.40) elle-même, malgré le fait que les régresseurs aient été multipliés par  $\mathbf{P}_W$ , parce que sous l'hypothèse nulle, la GNR n'aurait pas de pouvoir explicatif asymptotiquement. De ce fait, que l'on utilise la somme des carrés des résidus ou la somme des carrés expliquée, cela ne donne aucune différence asymptotiquement si nous voulons simplement tester l'hypothèse nulle  $\beta_2 = \mathbf{0}$ . Cela implique qu'un test en  $F$  ordinaire pour  $\mathbf{b}_2 = \mathbf{0}$  basé sur l'estimation OLS de (7.40) serait asymptotiquement valable.

Dans le cas des modèles de régression linéaire et non linéaire estimés par moindres carrés, il est possible de tester les hypothèses concernant  $\beta$  en utilisant des tests en  $F$  exacts ou asymptotiques de la forme

$$\frac{(\text{RSSR} - \text{USSR})/r}{\text{USSR}/(n - k)} \stackrel{a}{\sim} F(r, n - k), \quad (7.42)$$

où RSSR et USSR désignent la somme des carrés des résidus contraints et non contraints. Des tests de ce type sont également disponibles pour des modèles estimés par IV, mais ils ne sont pas véritablement les mêmes que (7.42), à

moins que, comme plus haut, RSSR et USSR ne soient obtenues par une GNR. A présent, nous allons discuter de ces tests plus en détails.

Pour simplifier, nous commençons par considérer le cas linéaire. Supposons que les modèles contraint et non contraint soient:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}), \quad \text{et} \quad (7.43)$$

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2\mathbf{I}), \quad (7.44)$$

et qu'ils soient estimés par IV, en utilisant la matrice d'instruments  $\mathbf{W}$ . Supposons maintenant que les estimations sont effectivement obtenues par les doubles moindres carrés. Il est facile de voir que la somme des carrés des résidus de la régression dans la seconde étape pour (7.43), dans laquelle  $\mathbf{X}_1$  est remplacée par  $\mathbf{P}_W\mathbf{X}_1$ , sera

$$\text{RSSR}^* \equiv \mathbf{y}^\top \mathbf{M}_1 \mathbf{y}, \quad (7.45)$$

où  $\mathbf{M}_1$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{P}_W\mathbf{X}_1)$ . De manière similaire, nous pouvons montrer (le faire constitue un bon exercice) que la somme des carrés des résidus de la régression de la seconde étape pour (7.44) sera

$$\text{USSR}^* \equiv \mathbf{y}^\top \mathbf{M}_1 \mathbf{y} - \mathbf{y}^\top \mathbf{M}_1 \mathbf{P}_W \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{P}_W \mathbf{M}_1 \mathbf{P}_W \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{P}_W \mathbf{M}_1 \mathbf{y}. \quad (7.46)$$

La différence entre (7.45) et (7.46) est

$$\mathbf{y}^\top \mathbf{M}_1 \mathbf{P}_W \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{P}_W \mathbf{M}_1 \mathbf{P}_W \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{P}_W \mathbf{M}_1 \mathbf{y}, \quad (7.47)$$

qui présente une ressemblance frappante mais en aucun cas innocente avec l'expression (7.41). Sous l'hypothèse nulle (7.43),  $\mathbf{y}$  est égal à  $\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}$ . Puisque  $\mathbf{P}_W\mathbf{M}_1$  annule  $\mathbf{X}_1$ , (7.47) se réduit à

$$\mathbf{u}^\top \mathbf{M}_1 \mathbf{P}_W \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{P}_W \mathbf{M}_1 \mathbf{P}_W \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{P}_W \mathbf{M}_1 \mathbf{u}$$

sous l'hypothèse nulle. Il serait facile de voir que, sous des hypothèses raisonnables, cette quantité divisée par une estimation convergente de  $\sigma^2$ , sera distribuée asymptotiquement par une  $\chi^2(r)$ . Les hypothèses nécessaires sont essentiellement (7.18a)–(7.18c), plus les hypothèses suffisantes pour qu'un théorème de la limite centrale puisse s'appliquer à  $n^{-1/2}\mathbf{W}^\top\mathbf{u}$ .

Alors, le problème est d'estimer  $\sigma^2$ . Remarquons que  $\text{USSR}^*/(n-k)$  n'estime pas  $\sigma^2$  de manière convergente, pour les raisons discutées à la Section 7.5. Comme nous l'avons vu, les résidus de la régression de la seconde étape seront trop grands, au moins asymptotiquement. Par conséquent, les estimations de  $\sigma^2$  doivent être basées sur l'ensemble des résidus  $\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$  plutôt que sur l'ensemble  $\mathbf{y} - \mathbf{P}_W\mathbf{X}\tilde{\boldsymbol{\beta}}$ . Une estimation valable est  $\text{USSR}/(n-k)$ , où

$$\text{USSR} \equiv \|\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1 - \mathbf{X}_2\tilde{\boldsymbol{\beta}}_2\|^2.$$



Une statistique de test analogue à (7.42) serait alors

$$\frac{(\text{RSSR}^* - \text{USSR}^*)/r}{\text{USSR}/(n-k)} \stackrel{a}{\sim} F(r, n-k). \quad (7.48)$$

Remarquons que le numérateur et le dénominateur de cette statistique de test sont basés sur les différents ensembles de résidus. Le numérateur est  $1/r$  fois la différence entre les sommes des carrés des résidus des régressions de la seconde étape, alors que le dénominateur est  $1/(n-k)$  fois la somme des carrés des résidus qui serait donnée par un programme pour l'estimation IV.

Malheureusement, très peu de progiciels de régression arrivent à obtenir en même temps des variantes des sommes des carrés des résidus, c'est-à-dire avec ou sans astérisque. Cela signifie que calculer une statistique de test comme (7.48) se révèle souvent plus difficile que cela ne le laisse paraître. Si nous utilisons une procédure d'estimation IV (ou 2SLS), les progiciels nous donneront, normalement, seulement la variante sans astérisque de la somme des carrés des résidus. Pour en obtenir la variante avec astérisque, nous devons alors retourner et exécuter la régression de la seconde étape par OLS. Souvenons-nous que l'on doit utiliser  $\text{RSSR}^* - \text{USSR}^*$  au numérateur de la statistique de test plutôt que  $\text{RSSR} - \text{USSR}$ , puisque seulement le premier est égal à (7.41). Pour une discussion plus détaillée, voir Startz (1983) et Wooldridge (1990c).

Nous retournons à présent au cas non linéaire. Que le modèle soit linéaire ou non, nous pouvons toujours utiliser des tests basés sur la valeur de la fonction critère (7.15). Pour les modèles non linéaires, il est naturel de construire un test sur la différence

$$\|\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}}))\|^2 - \|\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}}))\|^2. \quad (7.49)$$

Cette différence s'avère être asymptotiquement la même que la somme des carrés expliquée de la GNR (7.38), c'est-à-dire l'expression (7.39). Par conséquent, (7.49) divisée par une estimation convergente de  $\sigma^2$ , sera distribuée asymptotiquement suivant une  $\chi^2(r)$  sous l'hypothèse nulle  $\boldsymbol{\beta}_2 = \mathbf{0}$ . Ce résultat important sera démontré par la suite. Remarquons que la différence entre les deux valeurs de la fonction critère (7.15) n'est pas asymptotiquement la même que la somme des carrés expliquée de la régression de Gauss-Newton. Le résultat est exact dans ce cas particulier parce que les deux valeurs de la fonction critère correspondent aux valeurs contrainte et non contrainte de  $\boldsymbol{\beta}$ , et la GNR correspondant à la régression non contrainte est évaluée avec les valeurs contraintes.

Nous démontrons à présent ce résultat. A partir de (7.40), et du fait que  $\mathbf{P}_W$  est une matrice de projection, nous voyons que la somme des carrés expliquée et les estimations du paramètre  $\tilde{\mathbf{b}}$  de la GNR sont identiques à celles de la régression

$$\mathbf{P}_W(\mathbf{y} - \tilde{\mathbf{x}}) = \mathbf{P}_W\tilde{\mathbf{X}}_1\mathbf{b}_1 + \mathbf{P}_W\tilde{\mathbf{X}}_2\mathbf{b}_2 + \text{résidus}. \quad (7.50)$$

Supposons maintenant que dans la régression ci-dessus la contrainte  $\mathbf{b}_2 = \mathbf{0}$  soit imposée. Le résultat est

$$\mathbf{P}_W(\mathbf{y} - \tilde{\mathbf{x}}) = \mathbf{P}_W\tilde{\mathbf{X}}_1\mathbf{b}_1 + \text{résidus}. \quad (7.51)$$

La différence entre les sommes des carrés expliquées des régressions (7.50) et (7.51) est le numérateur de toutes les statistiques de test pour  $\beta_2 = \mathbf{0}$  qui sont basées sur la GNR (7.38). Cette différence est égale à la valeur absolue de la différence entre les deux sommes des carrés des résidus. La ESS de la régression (7.51) est égale à zéro, d'après les conditions du premier ordre pour la minimisation de la fonction somme-des-carrés contrainte. Donc, la SSR est juste la somme des carrés totale, c'est-à-dire

$$\|\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\check{\beta}))\|^2,$$

qui constitue le premier terme de l'expression (7.49).

Nous devons donc montrer à présent que la SSR de la régression (7.50) est asymptotiquement égale à l'opposé du second terme dans l'expression (7.49). Cette SSR est

$$\|\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\check{\beta}) - \tilde{\mathbf{X}}\check{\mathbf{b}})\|^2,$$

où  $\check{\mathbf{b}}$  est le vecteur des paramètres estimés provenant de l'estimation OLS de (7.50). Rappelons, à partir des résultats de la Section 6.6 sur l'estimation en une étape, que le vecteur  $(\tilde{\beta} - \check{\beta})$  est asymptotiquement égal à l'estimation  $\tilde{\mathbf{b}}$  de la GNR (7.38). Par conséquent,

$$\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\check{\beta}) - \tilde{\mathbf{X}}\check{\mathbf{b}}) \stackrel{a}{=} \mathbf{P}_W\mathbf{y} - \mathbf{P}_W\mathbf{x}(\check{\beta}) - \mathbf{P}_W\tilde{\mathbf{X}}(\tilde{\beta} - \check{\beta}). \quad (7.52)$$

Mais un développement de Taylor à l'ordre un de  $\mathbf{x}(\tilde{\beta})$  au point  $\beta = \check{\beta}$  nous donne

$$\mathbf{x}(\tilde{\beta}) \cong \mathbf{x}(\check{\beta}) + \mathbf{X}(\check{\beta})(\tilde{\beta} - \check{\beta}).$$

En soustrayant le côté droit de cette expression à  $\mathbf{y}$  et en multipliant le tout par  $\mathbf{P}_W$ , nous obtenons le côté droit de l'expression (7.52). Par conséquent, nous voyons que la SSR de la régression (7.50) est asymptotiquement égale à

$$\|\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\tilde{\beta}))\|^2,$$

qui constitue le second terme de (7.49). Nous avons donc démontré que la différence entre les valeurs contrainte et non contrainte de la fonction critère, l'expression (7.49), est asymptotiquement équivalente à la somme des carrés expliquée de la GNR (7.38). Puisque cette dernière peut être utilisée pour construire une statistique de test valable, la première le peut aussi.

Ce résultat est important. Il nous enseigne que nous pouvons toujours construire un test d'hypothèse sur  $\beta$  en prenant la différence entre les valeurs contrainte et non contrainte de la fonction critère pour l'estimation IV et en

la divisant par une estimation convergente de  $\sigma^2$ . De plus, un tel test serait asymptotiquement équivalent à prendre la somme des carrés expliquée de la GNR évaluée en  $\tilde{\beta}$  et à la traiter de la même manière. Chacun de ces tests peut être transformé en un test en  $F$  asymptotique en divisant le numérateur et le dénominateur par leurs degrés de liberté respectifs, c'est-à-dire  $r$  et  $n - k + r$ . C'est une bonne procédure à employer asymptotiquement, mais il n'est pas évident qu'elle soit avantageuse avec des échantillons finis.

## 7.8 IDENTIFICATION ET CONTRAINTES DE SURIDENTIFICATION

L'identification constitue une question un peu plus compliquée pour les modèles estimés par IV que pour les modèles estimés par moindres carrés, parce que le choix des instruments affecte l'identification du modèle. Un modèle qui ne serait pas identifié s'il était estimé par moindres carrés ne sera toujours pas identifié s'il est estimé par IV. Cependant, un modèle qui serait identifié s'il était estimé par moindres carrés peut ne pas être identifié s'il est estimé par IV selon le choix de la matrice des instruments. Cela se produira inévitablement s'il existe moins d'instruments que de paramètres. Cela peut se produire aussi bien dans d'autres circonstances.

Les conditions qui font qu'un modèle estimé par IV est identifié par un ensemble d'observations donné sont très similaires aux conditions rencontrées précédemment pour l'estimation NLS. Pour rester tout à fait dans un contexte général, supposons que l'on traite le modèle non linéaire (7.31). La fonction critère à minimiser correspond alors à l'expression (7.32). Comme nous l'avons vu dans les Chapitres 2 et 5, il existe au moins trois concepts d'identification auxquels on peut s'intéresser. Pour qu'un modèle soit **localement identifié** en un minimum local  $\tilde{\beta}$  de la fonction critère, la matrice des dérivées secondes de cette fonction doit être définie positive dans le voisinage de  $\tilde{\beta}$ . Pour un modèle **globalement identifié**, le minimum local  $\tilde{\beta}$  doit être l'unique minimum global de la fonction critère. Pour un modèle linéaire, l'identification locale implique l'identification globale, mais pour les modèles non linéaires ce n'est pas le cas. Le troisième concept d'identification est **l'identification asymptotique**. Pour qu'un modèle soit asymptotiquement identifié dans le voisinage de  $\beta_0$ ,

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top(\beta_0) \mathbf{P}_W \mathbf{X}(\beta_0))$$

doit exister et être une matrice définie positive quand la limite en probabilité est calculée sous n'importe quel DGP caractérisé par le vecteur paramétrique  $\beta_0$ . Aucune de ces conditions ne diffère d'une manière substantielle des conditions correspondantes pour les modèles de régression estimés par NLS. Les seules différences résultent, de manière évidente, de la présence de  $\mathbf{P}_W$  dans la fonction critère.

Les questions soulevées sur l'identification par l'utilisation des instruments sont les mêmes que le modèle soit linéaire ou non. Puisque les modèles

linéaires sont plus faciles à traiter, nous supposons pour la suite de cette section que le modèle est linéaire. Par conséquent, nous utiliserons le modèle (7.01). Il peut être estimé de manière convergente, soit en minimisant la fonction critère (7.15), soit par une procédure de doubles moindres carrés, à condition que la matrice d'instruments  $\mathbf{W}$  soit choisie de manière appropriée. Ces deux procédures donneront des estimations  $\tilde{\beta}$  identiques.

Il doit être évident que le modèle (7.01) ne sera ni localement ni globalement identifié si la matrice  $\mathbf{X}^\top \mathbf{P}_W \mathbf{X}$  n'est pas définie positive. Une condition nécessaire pour cela est que  $\rho(\mathbf{W}) \geq k$ . Normalement  $\rho(\mathbf{W})$  sera égal à  $l$ , le nombre d'instruments, et nous supposerons que c'est le cas. Par conséquent, la condition nécessaire exige qu'il y ait au moins autant d'instruments que de régresseurs. Cela n'est pourtant pas une condition suffisante. Si une combinaison linéaire des colonnes de  $\mathbf{X}$  se trouvait dans  $\mathcal{S}^\perp(\mathbf{W})$ ,  $\mathbf{P}_W \mathbf{X}$  serait de rang inférieur à  $k$  et  $\mathbf{X}^\top \mathbf{P}_W \mathbf{X}$  serait alors singulière, et donc ce cas doit être écarté. Lorsque  $\mathbf{X}^\top \mathbf{P}_W \mathbf{X}$  n'est pas singulière, le modèle (7.01) sera localement identifié et, parce qu'il est linéaire, il sera également identifié globalement. Il est souvent utile de distinguer deux types d'identifications locales. Lorsqu'il existe autant d'instruments que de régresseurs et que le modèle est identifié, le modèle est dit **juste identifié** ou **identifié exactement**, parce que la suppression d'une colonne quelconque de  $\mathbf{W}$  le rendrait non identifié. Si au contraire, il existe plus d'instruments que de régresseurs, si bien que le modèle resterait identifié si une (et peut-être plus d'une) colonne de  $\mathbf{W}$  était supprimée, le modèle sera dit **suridentifié**.

Les modèles linéaires qui sont exactement identifiés possèdent plusieurs propriétés intéressantes. Nous avons déjà vu que, pour un modèle identifié exactement, l'estimateur des IV généralisé (7.17) est égal à l'estimateur IV simple (7.25). Nous avons vu aussi que pour de tels modèles, l'estimateur IV n'aura pas de moments d'ordre supérieur ou égal à un. Une troisième propriété intéressante est que la valeur minimisée de la fonction critère (7.15) est exactement zéro. Comme nous le verrons plus loin, cette propriété est pratique pour certains usages.

Le résultat selon lequel la valeur minimisée de la fonction critère (7.15) est nulle quand  $l = k$  est important et révélateur. Cette valeur est

$$(\mathbf{y} - \mathbf{X}\tilde{\beta})^\top \mathbf{P}_W (\mathbf{y} - \mathbf{X}\tilde{\beta}) = (\mathbf{y} - \mathbf{X}\tilde{\beta})^\top \mathbf{P}_W \mathbf{y}.$$

Cette égalité résulte des conditions du premier ordre pour  $\tilde{\beta}$ , qui sont  $(\mathbf{y} - \mathbf{X}\tilde{\beta})^\top \mathbf{P}_W \mathbf{X} = \mathbf{0}$ . En utilisant l'expression (7.17), la valeur minimisée de la fonction critère peut alors se récrire comme

$$\mathbf{y}^\top \mathbf{P}_W \mathbf{y} - \mathbf{y}^\top \mathbf{P}_W \mathbf{X} (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y} = \mathbf{y}^\top (\mathbf{P}_W - \mathbf{P}_{\mathbf{P}_W \mathbf{X}}) \mathbf{y}. \quad (7.53)$$

La matrice située au milieu de l'expression de droite, est en réalité une matrice de projection orthogonale puisque  $\mathcal{S}(\mathbf{P}_W \mathbf{X}) \subseteq \mathcal{S}(\mathbf{W})$ . De plus, il est facile de voir que si  $\mathbf{W}^\top \mathbf{X}$  est une matrice carrée singulière,  $\mathcal{S}(\mathbf{P}_W \mathbf{X}) = \mathcal{S}(\mathbf{W})$ , et la

valeur de la fonction critère minimisée est donc nulle. L'intuition qui permet d'aboutir à ce résultat est très simple. Nous avons vu que l'estimation IV minimise seulement la portion de la distance entre  $\mathbf{y}$  et  $\mathcal{S}(\mathbf{X})$  qui appartient à  $\mathcal{S}(\mathbf{W})$ . Puisque dans ce cas  $\mathcal{S}(\mathbf{W})$  constitue un espace à  $k$  dimensions, et que nous minimisons par rapport aux  $k$  paramètres, la procédure de minimisation peut réduire cette distance à zéro, et par conséquent éliminer entièrement une quelconque disjonction entre  $\mathbf{P}_W \mathbf{y}$  et  $\mathcal{S}(\mathbf{P}_W \mathbf{X})$ .

Lorsque l'on dispose de plus d'instruments que de régresseurs (nous supposons encore que  $\rho(\mathbf{W})$  est égal à  $l$ ), le modèle sera dit suridentifié parce qu'il existe plus d'instruments qu'il n'en faut pour garantir l'identification. La terminologie vient de la littérature sur les modèles à équations simultanées dans lesquels l'identification a été étudiée en détails; voir le Chapitre 18. Nous avons vu dans la Section 7.5 qu'il est souvent préférable qu'un modèle soit quelque peu suridentifié afin de garantir de bonnes propriétés pour l'estimateur IV avec des échantillons finis. La possibilité de tester, dans une certaine mesure, la validité du choix des instruments constitue une seconde caractéristique attrayante des modèles suridentifiés. Cela est remarquablement facile à exécuter et peut être, dans certains cas, très instructif.

Quand nous spécifions un modèle comme (7.01), nous supposons que  $\mathbf{y}$  dépend linéairement de  $\mathbf{X}$  et ne dépend d'aucune autre variable observable. En particulier, nous supposons qu'il ne dépend pas des colonnes de  $\mathbf{W}$  qui n'appartiennent pas à  $\mathcal{S}(\mathbf{X})$ . Autrement, l'hypothèse d'indépendance entre les aléas  $\mathbf{u}$  de (7.01) et les instruments serait fausse. Dans certains cas, et peut-être même dans beaucoup de cas, nous ne pouvons pas être entièrement sûrs que les colonnes de  $\mathbf{W}$  puissent être écartées de  $\mathbf{X}$ . Néanmoins, nous devons exclure autant de colonnes de  $\mathbf{W}$  que de variables endogènes dans  $\mathbf{X}$  pour identifier le modèle. Et si le modèle est suridentifié, c'est que nous avons exclu encore plus de colonnes de  $\mathbf{W}$ . Ces contraintes supplémentaires, appelées **contraintes de suridentification**, peuvent être testées. Nous pouvons procéder par plusieurs moyens différents. Basmann (1960) est une référence classique, et d'autres encore comme Byron (1974), Wegge (1978), Hwang (1980), et Fisher (1981). Cependant, notre approche est beaucoup plus simple que celles évoquées dans ces articles.

Le moyen le plus simple de comprendre les tests des contraintes de suridentification consiste à les assimiler à des cas particuliers des tests d'hypothèses évoqués dans la section précédente. L'hypothèse nulle est le modèle (7.01). L'hypothèse alternative est donc le modèle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}^* \boldsymbol{\gamma} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (7.54)$$

où  $\mathbf{W}^*$  est une matrice composée de  $l - k$  colonnes de  $\mathbf{W}$  n'appartenant pas à  $\mathcal{S}(\mathbf{P}_W \mathbf{X})$ . Par conséquent, la matrice  $[\mathbf{P}_W \mathbf{X} \quad \mathbf{W}^*]$  de dimension  $n \times l$  sera de plein rang  $l$ . De plus,  $\mathcal{S}(\mathbf{P}_W \mathbf{X}, \mathbf{W}^*) = \mathcal{S}(\mathbf{W})$ . Comme nous le verrons dans un instant, pour calculer la statistique de test, il n'est pas vraiment nécessaire de construire  $\mathbf{W}^*$ .

Le modèle (7.54) est construit pour être juste identifié. Il y a, précisément, autant de régresseurs que d'instruments. Si (7.01) est spécifié correctement, dans une régression IV utilisant  $\mathbf{W}$  comme matrice d'instruments,  $\mathbf{W}^*$  ne doit pas avoir de pouvoir explicatif sur une quelconque variation de  $\mathbf{y}$  non expliquée par  $\mathbf{X}$ ;  $\gamma$  doit donc être égal à zéro. Si ce n'est pas le cas, c'est-à-dire que des colonnes de  $\mathbf{W}^*$  sont corrélées avec  $\mathbf{u}$ ,  $\gamma$  ne sera pas nul. Cela peut survenir soit parce que certaines colonnes de  $\mathbf{W}^*$  auraient dû être comprises dans  $\mathbf{X}$  alors qu'elles ne le sont pas, soit parce que certaines colonnes de  $\mathbf{W}^*$  sont corrélées avec  $\mathbf{u}$  et ne constituent pas, par conséquent, des instruments valables à utiliser. Donc, en testant l'hypothèse  $\gamma = \mathbf{0}$  nous pouvons tester l'hypothèse nulle *jointe* que (7.01) est spécifié correctement et que  $\mathbf{W}$  est une matrice d'instruments valable. Malheureusement, il est impossible de tester la dernière hypothèse de manière isolée.

Une fois le problème reformulé dans cette direction, il serait intéressant de pouvoir tester l'hypothèse  $\gamma = \mathbf{0}$  simplement en utilisant un des tests abordés dans la Section 7.7. Nous pourrions débiter par obtenir des estimations IV pour l'hypothèse nulle (7.01) et pour l'hypothèse alternative (7.54). Un test en  $F$  asymptotique pourrait alors être calculé comme

$$\frac{(\text{RSSR}^* - \text{USSR}^*)/(l - k)}{\text{USSR}/(n - l)} \underset{a}{\sim} F(l - k, n - l), \quad (7.55)$$

où  $\text{RSSR}^*$  et  $\text{USSR}^*$  sont les sommes des carrés des résidus OLS de la seconde étape à partir de l'estimation 2SLS de (7.01) et (7.54) respectivement, et  $\text{USSR}$  est la somme des carrés des résidus IV de (7.54).

Cependant, une autre procédure très simple est aussi acceptable. Nous avons vu pour le modèle non contraint (7.54) que la valeur de la fonction critère (7.15) doit être égale, à l'optimum, à zéro. Donc, la différence entre les valeurs contrainte et non contrainte de la fonction critère doit être égale à la valeur pour le modèle contraint. A partir de (7.53), la valeur de la fonction critère pour le modèle contraint est

$$\|P_W(\mathbf{y} - \mathbf{X}\tilde{\beta})\|^2 = \mathbf{y}^\top P_W \mathbf{y} - \mathbf{y}^\top P_W \mathbf{X} (\mathbf{X}^\top P_W \mathbf{X})^{-1} \mathbf{X}^\top P_W \mathbf{y}, \quad (7.56)$$

et il est facile de voir que (7.56) est égale à la différence entre  $\text{RSSR}^* - \text{USSR}^*$  qui apparaît dans (7.55). Cette expression peut être divisée par n'importe quelle quantité qui estime  $\sigma^2$  de façon convergente. Par conséquent, une statistique de test alternative est

$$\frac{\|P_W(\mathbf{y} - \mathbf{X}\tilde{\beta})\|^2}{\|(\mathbf{y} - \mathbf{X}\tilde{\beta})\|^2/n} \underset{a}{\sim} \chi^2(l - k), \quad (7.57)$$

expression qui correspond à  $n$  fois le  $R^2$  non centré de la régression des résidus IV  $\mathbf{y} - \mathbf{X}\tilde{\beta}$  sur les instruments  $\mathbf{W}$ . Il est alors facile de voir que cette régression

est équivalente à la GNR associée au modèle contraint (7.54). Par conséquent, nous voyons qu'il n'est pas nécessaire de spécifier  $\mathbf{W}^*$  ou d'estimer (7.54).

A partir de (7.57), nous voyons que si la valeur de la fonction critère est petite relativement à notre estimation de  $\sigma^2$ , nous ne pouvons pas rejeter l'hypothèse jointe où le modèle (7.01) serait correctement spécifié et où les instruments apparaissant dans  $\mathbf{W}$  seraient valides. Cela a un sens puisque la fonction critère (7.15) est la longueur au carré du vecteur  $\mathbf{P}_W(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ , qui constitue la contrepartie observable de  $\|\mathbf{P}_W\mathbf{u}\|^2$ . Si  $\mathbf{W}$  est une matrice valable d'instruments, les aléas  $\mathbf{u}$  ne devraient pas être corrélés avec  $\mathbf{W}$ , et de plus  $\|\mathbf{P}_W\mathbf{u}\|^2$  devrait être petit. Par conséquent, il est pertinent que le test soit basé sur  $\|\mathbf{P}_W(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\|^2$ . Cela a aussi un sens que le test possède  $l - k$  degrés de liberté puisque le numérateur ne sera pas égal à zéro dès que  $l > k$ .

La statistique de test (7.57) possède un analogue évident pour les modèles non linéaires comme (7.31). C'est

$$\frac{\|\mathbf{P}_W(\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}}))\|^2}{\|(\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}}))\|^2/n} \stackrel{a}{\sim} \chi^2(l - k), \quad (7.58)$$

et il peut se calculer en multipliant par  $n$  le  $R^2$  non centré d'une régression des résidus  $\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})$  sur la matrice d'instruments  $\mathbf{W}$ . En se basant sur les arguments exposés ci-dessus et sur ceux de la Section 7.7, les lecteurs peuvent vérifier que cette statistique de test est, en effet, distribuée asymptotiquement suivant une  $\chi^2(l - k)$  sous les conditions de régularité standard.

Les tests des contraintes de suridentification doivent être calculés automatiquement chaque fois que l'on procède à des estimations IV. Si, par hasard, la statistique de test est significativement plus grande que ce qu'elle doit être sous l'hypothèse nulle, on doit être extrêmement vigilant dans l'interprétation des estimations, puisqu'il est probable que le modèle soit mal spécifié ou que certains instruments ne soient pas valables. Chaque programme d'estimation IV de modèles de régression linéaire et non linéaire devrait calculer facilement les statistiques de test (7.57) et (7.58) automatiquement avec chaque ensemble de paramètres estimés pour les modèles suridentifiés. Malheureusement, à l'heure actuelle, beaucoup de programmes n'en sont pas capables.

## 7.9 LES TESTS DE DURBIN-WU-HAUSMAN

Jusqu'ici, nous avons supposé que l'investigateur sait toujours s'il est nécessaire d'utiliser les IV plutôt que les moindres carrés. Mais ce n'est pas toujours le cas. Parfois, la théorie économique suggère que certaines variables économiques pourraient être endogènes mais sans indiquer précisément si elles le sont, ou même si leur corrélation avec les aléas est telle que l'utilisation des moindres carrés produira un biais sérieux. Pour décider s'il est nécessaire

d'utiliser les IV, on doit se demander si un ensemble d'estimations obtenu par moindres carrés est convergent ou non. Dans cette section, nous traiterons des tests qui peuvent être utilisés pour répondre à cette question.

La question de savoir si un ensemble d'estimations est convergent est différente de la question à laquelle doivent répondre les statistiques de test que nous avons essayé d'analyser jusqu'à présent. Ces tests analysent simplement si certaines contraintes sur les paramètres d'un modèle peuvent être vérifiées. Par contraste, nous voulons à présent savoir si les paramètres qui nous intéressent dans un modèle ont été estimés de manière convergente. Dans un article très connu, Hausman (1978) proposa une famille de tests destinés à répondre à cette question. Son idée principale réside sur la possibilité de construire un test sur un **vecteur de contrastes**, c'est-à-dire, le vecteur de différence entre deux vecteurs d'estimations, où l'un serait convergent sous des conditions plus faibles que l'autre. Cette idée trouve son origine dans un article connu de Durbin (1954). Un des tests proposés par Hausman pour tester la convergence des estimations par moindres carrés fut aussi proposée par Wu (1973). De plus, nous nous référerons à tous les tests de ce genre en tant que **tests de Durbin-Wu-Hausman**, ou **tests DWH**. Ces dernières années, ces tests ont suscité beaucoup d'intérêt et beaucoup de travaux; voir en particulier Holly (1982), Ruud (1984), et Davidson et MacKinnon (1989). Dans cette section, nous traiterons seulement des idées fondamentales et nous montrerons comment des procédures de ce type peuvent être utilisées pour tester la convergence des estimations par moindres carrés lorsque des variables explicatives peuvent être endogènes. Dans le Chapitre 11, nous discuterons d'autres applications des tests DWH.

Supposons pour commencer que le modèle concerné soit (7.01). Le principe des tests DWH suggère que nous devons comparer l'estimateur OLS

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

avec l'estimateur IV

$$\tilde{\beta} = (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y}.$$

L'estimateur OLS  $\hat{\beta}$  serait convergent pour  $\beta_0$  si les données étaient générées par un cas particulier du modèle (7.01) avec  $\beta = \beta_0$ , et si  $\mathbf{u}$  était asymptotiquement indépendant de  $\mathbf{X}$ . L'estimateur IV  $\tilde{\beta}$  sera convergent pour  $\beta_0$  à condition que la matrice d'instruments  $\mathbf{W}$  satisfasse les conditions (7.18). Par conséquent, les deux estimateurs auront la même limite en probabilité,  $\beta_0$ . D'autre part, si  $\mathbf{u}$  n'était pas asymptotiquement indépendant de  $\mathbf{X}$ ,  $\tilde{\beta}$  serait encore convergent alors que  $\hat{\beta}$  ne le serait pas.



Le test DWH est basé sur le vecteur de contrastes

$$\begin{aligned}
 \tilde{\beta} - \hat{\beta} &= (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
 &= (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \left( \mathbf{X}^\top \mathbf{P}_W \mathbf{y} - (\mathbf{X}^\top \mathbf{P}_W \mathbf{X}) (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \right) \\
 &= (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \left( \mathbf{X}^\top \mathbf{P}_W (\mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{y} \right) \\
 &= (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{M}_X \mathbf{y}.
 \end{aligned} \tag{7.59}$$

Nous pourrions construire une statistique de test du  $\chi^2$  basée directement sur ce vecteur par un moyen assez évident. La statistique de test serait une forme quadratique du vecteur (7.59), avec au milieu l'inverse (généralisée)<sup>3</sup> d'une estimation de la matrice de covariance. Mais comme nous allons le voir, on peut ne pas construire un vecteur de contrastes pour calculer cette statistique de test DWH. En fait, comme dans l'article de Davidson et MacKinnon (1989) le montre, on n'a jamais besoin de construire un vecteur de contrastes pour calculer une statistique DWH. Ces statistiques peuvent toujours être calculées au moyen de régressions artificielles.

Le premier facteur dans (7.59),  $(\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1}$ , correspond simplement à une matrice de dimension  $k \times k$  de plein rang. Evidemment, sa présence n'aura aucun effet sur une statistique de test que nous pourrions calculer. Par conséquent, ce que nous voulons vraiment faire, c'est tester si le vecteur

$$\mathbf{X}^\top \mathbf{P}_W \mathbf{M}_X \mathbf{y} \tag{7.60}$$

est d'espérance nulle asymptotiquement. Il doit l'être parce que sous un DGP appartenant à (7.01), il est égal à

$$\mathbf{X}^\top \mathbf{P}_W \mathbf{M}_X \mathbf{u}.$$

Ce vecteur possède  $k$  éléments, mais même si  $\mathbf{P}_W \mathbf{X}$  est de plein rang, tous ces éléments peuvent ne pas être des variables aléatoires parce que  $\mathbf{M}_X$  peut annuler des colonnes de  $\mathbf{P}_W \mathbf{X}$ . Supposons que  $k^*$  soit le nombre de colonnes linéairement indépendantes de  $\mathbf{P}_W \mathbf{X}$  qui ne sont pas annulées par  $\mathbf{M}_X$ . Alors, si l'on désigne par  $\mathbf{X}^*$  ces  $k^*$  colonnes, il serait intéressant de tester la nullité asymptotique du vecteur

$$\mathbf{X}^{*\top} \mathbf{P}_W \mathbf{M}_X \mathbf{y}. \tag{7.61}$$

<sup>3</sup> On doit utiliser une inverse généralisée dans les cas où la matrice de covariance du vecteur de contrastes n'est pas de plein rang. Voir Hausman et Taylor (1982).

A présent, considérons la régression artificielle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}' + \mathbf{P}_W\mathbf{X}^*\boldsymbol{\delta} + \text{résidus.} \quad (7.62)$$

Les régresseurs  $\mathbf{P}_W\mathbf{X}^*$  représentent les valeurs ajustées provenant de la régression de  $\mathbf{X}^*$ , les colonnes de  $\mathbf{X}$  n'appartenant pas à  $\mathcal{S}(\mathbf{W})$ , sur la matrice des instruments  $\mathbf{W}$ . Puisque

$$\mathcal{S}(\mathbf{X}, \mathbf{P}_W\mathbf{X}^*) = \mathcal{S}(\mathbf{X}, \mathbf{P}_W\mathbf{X}) = \mathcal{S}(\mathbf{X}, \mathbf{M}_W\mathbf{X}) = \mathcal{S}(\mathbf{X}, \mathbf{M}_W\mathbf{X}^*),$$

la régression (7.62) doit avoir la même SSR que la régression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}_W\mathbf{X}^*\boldsymbol{\eta} + \text{résidus,} \quad (7.63)$$

où les régresseurs  $\mathbf{M}_W\mathbf{X}^*$  sont les résidus de la régression de  $\mathbf{X}^*$  sur  $\mathbf{W}$ . Le test DWH peut être construit à partir de chacune de ces régressions. C'est simplement le test en  $F$  pour  $\boldsymbol{\delta} = \mathbf{0}$  dans (7.62) ou le test en  $F$  pour  $\boldsymbol{\eta} = \mathbf{0}$  dans (7.63). Du fait que (7.62) et (7.63) génèrent la même somme des carrés des résidus, il est clair que ces deux tests seront numériquement identiques.

Grâce au Théorème FWL, l'estimation OLS de  $\boldsymbol{\delta}$  dans (7.62) est

$$\tilde{\boldsymbol{\delta}} = (\mathbf{X}^{*\top}\mathbf{P}_W\mathbf{M}_X\mathbf{P}_W\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{P}_W\mathbf{M}_X\mathbf{y}.$$

En général, il est évident que  $\text{plim}(\tilde{\boldsymbol{\delta}}) = \mathbf{0}$  si et seulement si (7.61) a une espérance nulle asymptotiquement. Le  $F$  de Fisher ordinaire pour  $\boldsymbol{\delta} = \mathbf{0}$  dans (7.62) est

$$\frac{\mathbf{y}^\top\mathbf{P}_{\mathbf{M}_X\mathbf{P}_W\mathbf{X}^*}\mathbf{y}/k^*}{\mathbf{y}^\top\mathbf{M}_{\mathbf{X}, \mathbf{M}_X\mathbf{P}_W\mathbf{X}^*}\mathbf{y}/(n-k-k^*)} \stackrel{a}{\sim} F(k^*, n-k-k^*), \quad (7.64)$$

où  $\mathbf{P}_{\mathbf{M}_X\mathbf{P}_W\mathbf{X}^*}$  désigne la matrice qui projette orthogonalement sur l'espace  $\mathcal{S}(\mathbf{M}_X\mathbf{P}_W\mathbf{X}^*)$ , et  $\mathbf{M}_{\mathbf{X}, \mathbf{M}_X\mathbf{P}_W\mathbf{X}^*}$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}, \mathbf{M}_X\mathbf{P}_W\mathbf{X}^*)$ . Si (7.01) génère effectivement les données et que  $\mathbf{X}$  est indépendante de  $\mathbf{u}$ , la statistique (7.64) sera clairement valable asymptotiquement puisque le dénominateur estimera alors  $\sigma^2$  de manière convergente. Elle sera distribuée exactement suivant une  $F(k^*, n-k-k^*)$  avec des échantillons finis si les  $u_t$  dans (7.01) sont normalement distribués, et si  $\mathbf{X}$  et  $\mathbf{P}_W$  peuvent être considérés comme fixes.

Cette version du test DWH est souvent *interprétée* comme un test d'exogénéité des composantes de  $\mathbf{X}$  n'appartenant pas à l'espace engendré par  $\mathbf{W}$ ; voir Wu (1973), Hausman (1978), et Nakamura et Nakamura (1981). Cette interprétation est assez trompeuse puisque ce qui est testé n'est pas l'exogénéité ou l'endogénéité des composantes de  $\mathbf{X}$ , mais plutôt l'effet d'une éventuelle endogénéité sur les estimations de  $\boldsymbol{\beta}$ . L'hypothèse nulle est que les estimations OLS  $\hat{\boldsymbol{\beta}}$  sont convergentes et non pas que chaque colonne de  $\mathbf{X}$  est asymptotiquement indépendante de  $\mathbf{u}$ . Quoi qu'il en soit, cette version du test DWH peut se révéler utile lorsqu'il existe un doute sur la validité de l'usage des moindres carrés plutôt que des variables instrumentales.

La régression (7.63) mérite plus de commentaires. Elle possède la particularité remarquable que les estimations OLS et IV de  $\beta$  sont numériquement identiques dans le modèle (7.01). De plus, les matrices de covariance estimées sont aussi les mêmes, excepté que l'estimation OLS de (7.63) utilise un estimateur non convergent de  $\sigma^2$ . Ces résultats sont faciles à obtenir. Notons  $M^*$  la projection orthogonale sur l'espace  $\mathcal{S}^\perp(M_W X^*)$ . Grâce au Théorème FWL, les estimations OLS de (7.63) doivent être identiques à celles de la régression

$$M^* y = M^* X \beta + \text{residuals}. \quad (7.65)$$

Or,

$$M^* X = X - M_W X^* (X^{*\top} M_W X^*)^{-1} X^{*\top} M_W X.$$

Du fait que  $M_W X = [M_W X^* \ 0]$ , il s'ensuit que

$$X^{*\top} M_W X = X^{*\top} M_W [X^* \ 0].$$

Par conséquent, on a

$$M^* X = X - [M_W X^* \ 0] = X - M_W X = P_W X.$$

L'estimation OLS de  $\beta$  dans (7.65) devient

$$(X^\top M^* X)^{-1} X^\top M^* y = (X^\top P_W X)^{-1} X^\top P_W y. \quad (7.66)$$

Bien sûr, le membre de droite de (7.66) désigne l'expression de l'estimation IV ou 2SLS de  $\beta$ , notée (7.17).

En approfondissant cet argument, il est facile de voir que la matrice de covariance estimée par OLS de  $\hat{\beta}$  (7.63) sera

$$\tilde{s}^2 (X^\top P_W X)^{-1}, \quad (7.67)$$

où  $\tilde{s}^2$  désigne dans (7.63) l'estimation OLS de la variance des erreurs. L'expression (7.67) ressemble assez à la matrice de covariance IV (7.24), excepté que  $\tilde{s}^2$  apparaît à la place de  $\tilde{\sigma}^2$ . Lorsque  $\eta$  n'est pas nul (pour que l'estimation IV soit nécessaire), la variance des erreurs dans (7.63) sera plus petite que  $\sigma^2$ . Par conséquent,  $\tilde{s}^2$  sera biaisée vers le bas en tant qu'estimateur de  $\sigma^2$ . Bien sûr, il serait facile d'obtenir une matrice de covariance estimée valable en multipliant (7.67) par  $\tilde{\sigma}^2/\tilde{s}^2$ .

A présent, retournons au test DWH. Une variante de ce test s'applique aussi bien aux modèles linéaires qu'aux modèles non linéaires comme (7.31). Le test serait construit alors sur une variante de la régression de Gauss-Newton. Si l'hypothèse nulle était que les estimations NLS  $\hat{\beta}$  étaient convergentes, une statistique de test appropriée serait un test en  $F$  asymptotique pour  $c = 0$  dans la GNR

$$y - \hat{x} = \hat{X} b + M_W \hat{X}^* c + \text{résidus}, \quad (7.68)$$

où, comme d'habitude,  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\boldsymbol{\beta}})$  et  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\boldsymbol{\beta}})$ . C'est un bon exercice de vérifier que cette procédure donne une statistique de test qui est asymptotiquement équivalente à la statistique de test que l'on obtiendrait si l'on commençait avec le vecteur de contrastes  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ .

Il existe un problème avec les tests DWH pour les modèles non linéaires. Dans le cas d'un modèle non linéaire, il n'est pas toujours évident de savoir quelles colonnes de  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\boldsymbol{\beta}})$  doivent être comprises dans  $\hat{\mathbf{X}}^*$  et par conséquent, de déterminer le nombre de degrés de liberté du test. Asymptotiquement, nous voulons comprendre dans  $\mathbf{X}^*$  toutes les colonnes de  $\mathbf{X}$  qui n'appartiennent pas à  $\mathcal{S}(\mathbf{W})$ . Le nombre de degrés de liberté pour le test sera donc

$$\rho([\mathbf{X}(\boldsymbol{\beta}_0) \quad \mathbf{M}_W \mathbf{X}(\boldsymbol{\beta}_0)]) - k.$$

Le problème est que  $\mathbf{X}(\boldsymbol{\beta}_0)$  dépend de  $\boldsymbol{\beta}_0$ , qui, bien évidemment, est inconnu. En pratique, il est nécessaire d'utiliser  $\hat{\mathbf{X}}$  au lieu de  $\mathbf{X}(\boldsymbol{\beta}_0)$ . Malheureusement, il est possible que le rang de  $[\hat{\mathbf{X}} \quad \mathbf{M}_W \hat{\mathbf{X}}]$  ne soit pas le même que le rang de  $[\mathbf{X}(\boldsymbol{\beta}_0) \quad \mathbf{M}_W \mathbf{X}(\boldsymbol{\beta}_0)]$ . Quand cela se produit, la statistique de test calculée à partir de (7.68) aura le nombre de degrés de liberté erroné.

Quelquefois, il peut être difficile de décider des variables à utiliser en tant qu'instruments et celles qui doivent être traitées comme endogènes. Dans certains cas, il peut être utile de tester la convergence d'un estimateur IV en utilisant un test DWH. On le réalise facilement: supposons que l'on sache que les variables explicatives doivent être traitées comme endogènes mais qu'il ne soit pas évident que  $q$  autres puissent l'être. Quand le plus petit nombre doit être traité comme endogène, la matrice des instruments doit être  $\mathbf{W}_1$ , et quand le nombre le plus grand doit être traité comme endogène, elle doit être  $\mathbf{W}_2$ , où  $\mathbf{W}_1$  inclut toutes les colonnes qui sont dans  $\mathbf{W}_2$  plus  $q$  autres colonnes de  $\mathbf{X}$ . Considérons le cas linéaire. Les deux estimateurs sont

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_1 &= (\mathbf{X}^\top \mathbf{P}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_1 \mathbf{y} \quad \text{et} \\ \tilde{\boldsymbol{\beta}}_2 &= (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_2 \mathbf{y},\end{aligned}$$

où  $\mathbf{P}_1$  et  $\mathbf{P}_2$  sont, respectivement, les matrices qui projettent orthogonalement sur  $\mathcal{S}(\mathbf{W}_1)$  et  $\mathcal{S}(\mathbf{W}_2)$ . Le vecteur de contrastes est

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_2 - \tilde{\boldsymbol{\beta}}_1 &= (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_2 \mathbf{y} - (\mathbf{X}^\top \mathbf{P}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_1 \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1} \left( \mathbf{X}^\top \mathbf{P}_2 \mathbf{y} - (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X}) (\mathbf{X}^\top \mathbf{P}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_1 \mathbf{y} \right) \\ &= (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1} \left( \mathbf{X}^\top \mathbf{P}_2 (\mathbf{I} - \mathbf{P}_1 \mathbf{X} (\mathbf{X}^\top \mathbf{P}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_1) \mathbf{y} \right) \\ &= (\mathbf{X}^\top \mathbf{P}_2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_2 \mathbf{M}_{\mathbf{P}_1 \mathbf{X}} \mathbf{y},\end{aligned} \tag{7.69}$$

où  $\mathbf{M}_{\mathbf{P}_1 \mathbf{X}}$  est la matrice qui projette sur  $\mathcal{S}^\perp(\mathbf{P}_1 \mathbf{X})$ . Ici, la troisième ligne est déduite de l'égalité  $\mathbf{P}_2 \mathbf{P}_1 = \mathbf{P}_2$ , qui est une conséquence du fait que  $\mathcal{S}(\mathbf{W}_2)$  est un sous-espace de  $\mathcal{S}(\mathbf{W}_1)$ .

Nous laissons en tant qu'exercice la démonstration que le test de la nullité asymptotique de l'espérance du vecteur (7.69) peut être accompli en testant la nullité du vecteur  $\delta$  à  $q$  composantes dans la régression

$$y = X\beta + P_2 X^* \delta + \text{résidus.} \quad (7.70)$$

Ici  $P_2 X^*$  contient les  $q$  colonnes de  $P_2 X$  qui ne sont pas annulées par  $M_{P_1 X}$ . La régression (7.70) doit être estimée par IV en utilisant  $W_1$  comme matrice d'instruments, et un des tests discutés dans la Section 7.7 peut être utilisé pour tester  $\delta = 0$ .

## 7.10 CONCLUSION

Ce chapitre a introduit tous les concepts importants associés à la technique d'estimation des variables instrumentales. Pour un traitement plus détaillé, voir Bowden et Turkington (1984). Une autre référence utile est Godfrey (1988, Chapitre 5), où est abordé un nombre important de tests de spécification pour les modèles linéaires et non linéaires qui ont été estimés par IV.

Dans ce chapitre, nous avons appliqué la méthode des variables instrumentales seulement aux modèles de régression univariée linéaire et non linéaire avec les erreurs i.i.d. Nous rencontrerons plus tard dans cet ouvrage de nombreuses autres applications, notamment dans les Chapitres 17 et 18 où nous discutons, respectivement, de l'estimation GMM et des modèles à équations simultanées. Dans bien d'autres cas, nous exposerons un résultat dans le contexte d'une estimation OLS ou NLS, et nous montrerons qu'il tend vers une modification mineure dans le contexte d'une estimation IV.

## TERMES ET CONCEPTS

biais des équations simultanées	forme réduite non contrainte (URF)
contraintes de suridentification	forme structurelle (d'un modèle
erreurs dans les variables	d'équations simultanées)
estimateur doubles moindres carrés	identification: locale, globale, et
(2SLS)	asymptotique
estimateur des variables	instruments (variables instrumentales)
instrumentales (IV)	modèle d'équations simultanées
estimateur IV généralisé	modèle exactement identifié (ou juste
estimateur IV simple	identifié)
estimateur non linéaire des doubles	modèle suridentifié
moindres carrés (NL2SLS)	normalisation (d'un modèle
estimateur non linéaire IV	d'équations simultanées)
fonction critère	régression de Gauss-Newton (GNR)
forme réduite contrainte (RRF)	tests de Durbin-Wu-Hausman (DWH)
forme réduite (d'un modèle	variable prédéterminée
d'équations simultanées)	vecteur de contrastes

# Chapitre 8

## La Méthode du Maximum de Vraisemblance

### 8.1 INTRODUCTION

Les techniques d'estimation dont nous avons discuté jusqu'ici – moindres carrés et variables instrumentales – sont applicables uniquement aux modèles de régression. Mais tous les modèles ne peuvent pas s'écrire comme une égalité entre la variable dépendante et une fonction de régression plus un terme d'erreur, ou de telle sorte qu'un ensemble de variables dépendantes, sous la forme d'un vecteur, soit égal à un vecteur de fonctions de régression plus un vecteur d'aléas (Chapitre 9). Dans ces cas, les moindres carrés et les variables instrumentales ne sont tout simplement pas appropriés. Dans ce chapitre, nous introduisons par conséquent une troisième méthode d'estimation, qui est beaucoup plus largement applicable que les techniques dont nous avons discuté jusqu'ici, mais qui nécessite également d'assez fortes hypothèses. Il s'agit de l'estimation par la méthode du **maximum de vraisemblance**, ou **ML**.

A titre d'exemple du manque de pertinence des moindres carrés, considérons le modèle

$$y_t^\gamma = \beta_0 + \beta_1 x_t + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad (8.01)$$

qui ressemble presque à un modèle de régression. Ce modèle a du sens tant que le membre de droite de (8.01) demeure toujours positif, et il peut même être un modèle attrayant dans certains cas.<sup>1</sup> Par exemple, supposons que les observations portant sur  $y_t$  soient inclinées à droite mais que celles portant sur  $x_t$  ne le soient pas. Alors un modèle de régression conventionnel pourrait réconcilier ces deux faits uniquement si les aléas  $u_t$  étaient inclinés à droite, ce que l'on ne voudrait probablement pas supposer et qui rendrait l'utilisation des moindres carrés douteuse. D'un autre côté, le modèle (8.01) avec  $\gamma < 1$

<sup>1</sup> A proprement parler, il est impossible, naturellement, de garantir que le membre de droite de (8.01) soit toujours positif, mais ce modèle peut être considéré comme une très bonne approximation si  $\beta_0 + \beta_1 x_t$  est toujours plus grand que  $\sigma$ .

pourrait bien être capable de réconcilier ces faits tout en permettant aux aléas d'avoir une distribution symétrique.

Si  $\gamma$  était connu, (8.01) serait un modèle de régression. Mais si  $\gamma$  doit être estimé, (8.01) *n'est pas* un modèle de régression. Par conséquent, il ne peut pas être raisonnablement estimé par moindres carrés. La fonction somme-des-carrés est

$$SSR(\beta, \gamma) = \sum_{t=1}^n (y_t^\gamma - \beta_0 - \beta_1 x_t)^2,$$

et si, par exemple, tous les  $y_t$  étaient plus grands que l'unité, il est clair que cette fonction pourrait être arbitrairement construite proche de zéro simplement en laissant tendre  $\gamma$  vers moins l'infini et en posant  $\beta_0$  et  $\beta_1$  égaux à zéro. Par conséquent, personne ne pourrait *jamaïs* obtenir des estimations sensées de (8.01) en utilisant les moindres carrés ordinaires. Cependant, ce modèle peut être estimé très facilement en utilisant la méthode du maximum de vraisemblance qui sera expliquée dans la Section 8.10.

L'idée fondamentale de l'estimation par maximum de vraisemblance est, comme le nom l'implique, de trouver un ensemble d'estimations de paramètres, appelé  $\hat{\theta}$ , telles que la **vraisemblance** d'avoir obtenu l'échantillon que nous utilisons soit maximisée. Nous signifions par là que la densité de probabilité jointe pour le modèle que l'on estime est évaluée aux valeurs observées de la (des) variable(s) dépendante(s) et traitée comme une fonction de paramètres du modèle. Le vecteur  $\hat{\theta}$  des estimations ML donne alors le maximum de cette fonction. Ce principe d'estimation est très largement applicable: si nous pouvons écrire la densité jointe de l'échantillon, nous pouvons en principe utiliser le maximum de vraisemblance, soumis bien sûr à certaines conditions de régularité. Par ailleurs, il a un nombre de propriétés extrêmement commodes, dont nous discuterons brièvement dans ce qui suit et plus en détail dans le reste de ce chapitre. Il possède également quelques propriétés peu pratiques, et pour cela, le praticien doit parfois être méfiant.

La manière la plus simple de saisir l'idée fondamentale de l'estimation par ML est de considérer un exemple simple. Supposons que chaque observation  $y_t$  soit générée par la densité

$$f(y_t, \theta) = \theta e^{-\theta y_t}, \quad y_t > 0, \quad \theta > 0, \quad (8.02)$$

et soit indépendante de toutes les autres  $y_t$ . Il s'agit de la densité de la **distribution exponentielle**.<sup>2</sup> Il y a un seul paramètre inconnu  $\theta$  que nous

<sup>2</sup> La distribution exponentielle est utile pour l'analyse des phénomènes tels que les files d'attente ou les durées du chômage. Consulter n'importe quel ouvrage de statistique de niveau avancé, tel que Cox et Hinkley (1974) ou Hogg et Craig (1978). Pour des traitements plus précis, consulter, entre autres, Cox et Oakes (1984), Lawless (1982), et Miller (1981). Voir Kiefer (1988) et Lancaster (1990) pour des applications économiques.



désirons estimer, et nous disposons de  $n$  observations avec lesquelles nous allons travailler. La densité jointe des  $y_t$  sera désignée sous le nom de **fonction de vraisemblance** et notée  $L(\mathbf{y}, \theta)$ ; pour toute valeur de  $\theta$ , cette fonction nous renseigne sur la probabilité que nous aurions eue d'observer l'échantillon  $\mathbf{y} \equiv [y_1 \vdots \cdots \vdots y_n]$ .

Comme les  $y_t$  sont indépendants, leur densité jointe est simplement le produit de leurs densités marginales. Ainsi, la fonction de vraisemblance s'écrit

$$L(\mathbf{y}, \theta) = \prod_{t=1}^n \theta e^{-\theta y_t}. \quad (8.03)$$

Dans le cas d'échantillons de grande taille, (8.03) peut devenir extrêmement importante ou extrêmement petite, et prendre des valeurs qui sont bien au-delà des possibilités des nombres à virgule flottante que les ordinateurs manipulent. Pour cette raison, parmi d'autres, il est d'usage de maximiser le *logarithme* de la fonction de vraisemblance plutôt que la fonction de vraisemblance elle-même. Bien évidemment, nous obtiendrons la même réponse en procédant ainsi, car la **fonction de logvraisemblance**  $\ell(\mathbf{y}, \theta) \equiv \log(L(\mathbf{y}, \theta))$  est une fonction monotone croissante de  $L(\mathbf{y}, \theta)$ ; si  $\hat{\theta}$  maximise  $\ell(\mathbf{y}, \theta)$ , il doit aussi maximiser  $L(\mathbf{y}, \theta)$ . Dans le cas de (8.03), la fonction de logvraisemblance est

$$\ell(\mathbf{y}, \theta) = \sum_{t=1}^n (\log(\theta) - \theta y_t) = n \log(\theta) - \theta \sum_{t=1}^n y_t. \quad (8.04)$$

La maximisation de la fonction de logvraisemblance par rapport au seul paramètre inconnu  $\theta$ , est une procédure directe. Différentier l'expression la plus à droite de (8.04) par rapport à  $\theta$  et poser la dérivée à zéro donne la condition du premier ordre

$$\frac{n}{\theta} - \sum_{t=1}^n y_t = 0, \quad (8.05)$$

et nous trouvons pour la résolution de l'estimateur ML  $\hat{\theta}$  que

$$\hat{\theta} = \frac{n}{\sum_{t=1}^n y_t}. \quad (8.06)$$

Dans ce cas, il n'est pas nécessaire de se soucier des multiples solutions de (8.05). La dérivée seconde de (8.04) est toujours négative, ce qui nous permet de conclure que  $\hat{\theta}$  défini par (8.06) est *l'unique* estimateur ML. Notons que cela ne sera pas toujours le cas; pour certains problèmes les conditions du premier ordre peuvent mener à des solutions multiples.

Dès à présent, nous pourrions à juste titre poser certaines questions relatives aux propriétés de  $\hat{\theta}$ . Est-ce dans tous les sens du terme un bon estimateur à utiliser? Est-il biaisé? Est-il convergent? Comment est-il distribué? Et

ainsi de suite. Nous pourrions certainement étudier ces questions pour ce cas particulier. Mais une grande part de cette investigation se révélerait inutile, car le fait que  $\hat{\theta}$  soit un estimateur ML nous renseigne immédiatement sur un grand nombre de ses propriétés. C'est, en effet, une des caractéristiques les plus attrayantes de l'estimation ML: parce que beaucoup d'éléments sur les propriétés des estimateurs ML sont généralement connus, nous n'avons pas toujours besoin de pratiquer une étude particulière dans tous les cas.

Deux propriétés attrayantes majeures des estimateurs ML sont la **convergence** et la **normalité asymptotique**. Celles-ci sont des propriétés que nous avons déjà longuement étudiées dans le contexte des moindres carrés, et à ce titre nous n'avons pas besoin de les présenter davantage. Une troisième propriété attrayante est l'**efficacité asymptotique**. Ceci est vrai dans un sens plus fort pour les estimateurs ML que pour ceux des moindres carrés; comme nous n'avons pas formulé de fortes hypothèses sur la distribution des aléas lorsque nous discutons des moindres carrés, nous ne pouvions qu'affirmer que les estimations par moindres carrés non linéaires étaient asymptotiquement efficaces à l'intérieur d'une classe d'estimateurs assez limitée. Comme la méthode du maximum de vraisemblance nous force à expliciter en partie les hypothèses de distribution des aléas, nous serons capables de prouver des résultats plus forts.

Le fait que la matrice de covariance des estimations des paramètres résultant de l'estimation par ML puisse être estimée sans difficulté de différentes façons est étroitement lié à ces propriétés. Plus loin, comme nous le verrons dans la Section 8.9, la procédure ML conduit naturellement à plusieurs statistiques de test asymptotiquement équivalentes, dont au moins une d'entre elles peut être calculée aisément. Les estimations ML en elles-mêmes sont directement calculables, parce que la maximisation, même la maximisation non linéaire, est une procédure très bien comprise et, au moins conceptuellement, facile à effectuer. Ainsi une des qualités les plus appréciables de l'estimateur ML est son **calcul**: les estimations ML, aussi bien que les écarts types estimés et les statistiques de test, peuvent généralement être calculés de manière directe, bien que parfois coûteuse.

Une cinquième propriété souhaitable des estimateurs ML est l'**invariance**, terme par lequel nous signifions l'invariance à la reparamétrisation du modèle. Ceci est facile à illustrer à travers l'exemple que nous considérons jusqu'ici. Supposons que nous ayons paramétrisé la densité de  $y_t$  comme

$$f'(y_t, \phi) = (1/\phi)e^{-y_t/\phi}, \quad (8.07)$$

où  $\phi \equiv 1/\theta$ . Il est facile de décrire la relation entre  $\hat{\phi}$  et  $\hat{\theta}$ . La logvraisemblance dans la paramétrisation en  $\phi$  est

$$\ell'(\mathbf{y}, \phi) = \sum_{t=1}^n \left( -\log(\phi) - \frac{y_t}{\phi} \right) = -n \log(\phi) - \frac{1}{\phi} \sum_{t=1}^n y_t.$$

La condition de premier ordre pour un maximum de  $\ell'$  est alors

$$-\frac{n}{\phi} + \frac{1}{\phi^2} \sum_{t=1}^n y_t = 0,$$

et l'estimation ML décrite par  $\hat{\phi}$  est donc

$$\hat{\phi} = \frac{1}{n} \sum_{t=1}^n y_t = \frac{1}{\hat{\theta}}.$$

Nous constatons que la relation entre  $\hat{\phi}$  et  $\hat{\theta}$  est exactement la même que celle établie entre  $\phi$  et  $\theta$ . Alors, dans ce cas, l'estimation ML est **invariante à la reparamétrisation**. En fait, ceci est une propriété générale du maximum de vraisemblance. Tout spécialement dans les cas où la reparamétrisation est plus ou moins arbitraire, elle peut être une de ses propriétés les plus attrayantes.

Les propriétés du ML ne sont pas toutes enviables. Une caractéristique indésirable majeure concerne la dépendance aux hypothèses explicites de distribution des aléas, que le chercheur ressent souvent comme étant trop forte. Ceci n'est pas toujours un problème aussi sérieux que ce qu'il peut paraître. Bien qu'*en général* les propriétés asymptotiques des estimateurs ML soient valables seulement lorsque le modèle est correctement spécifié à tous les égards, nombreux sont les cas où une ou plusieurs de ces propriétés restent valides malgré quelques spécifications douteuses. Par exemple, l'estimateur des moindres carrés non linéaires correspond à l'estimateur par maximum de vraisemblance lorsque le modèle est un modèle de régression non linéaire à aléas normaux et indépendants (consulter la Section 8.10) et, comme nous l'avons vu, la convergence et la normalité asymptotique des NLS ne nécessitent pas l'hypothèse de normalité des aléas. Ainsi lorsque les aléas ne sont pas normaux, l'estimateur des moindres carrés non linéaires est un exemple de l'**estimateur quasi-ML**, ou **estimateur QML**, c'est-à-dire un estimateur ML appliqué à une situation pour laquelle il n'est pas entièrement valable; voir White (1982) et Gouriéroux, Monfort, Trognon (1984). Les estimateurs QML sont aussi parfois appelés **estimateurs pseudo-ML**.

L'autre caractéristique majeure indésirable du ML est que ses propriétés avec des échantillons finis peuvent être très différentes de ces propriétés asymptotiques. Bien qu'elles soient convergentes, les estimations des paramètres ML sont typiquement biaisées, et les estimations de la matrice de covariance ML peuvent être sérieusement trompeuses. Parce qu'en pratique les propriétés avec des échantillons finis sont souvent inconnues, le chercheur doit décider (souvent sans beaucoup d'information) comment se fier aux propriétés asymptotiques connues. Ceci introduit un facteur d'imprécision dans les efforts fournis pour établir des inférences par ML quand la taille de l'échantillon n'est pas extrêmement importante.

Dans le reste de ce chapitre, nous discuterons des propriétés les plus importantes du maximum de vraisemblance. La relation entre les moindres carrés et le maximum de vraisemblance sera introduite à la Section 8.10 et sera aussi un des thèmes abordés dans le Chapitre 9, qui s'intéresse principalement aux moindres carrés généralisés et à leur relation avec le ML. Des exemples d'estimation par maximum de vraisemblance en économétrie seront fournis dans la suite du livre. Des exemples complémentaires peuvent être trouvés chez Cramer (1986).

## 8.2 CONCEPTS FONDAMENTAUX ET NOTATION

L'estimation par maximum de vraisemblance repose sur la notion de **vraisemblance** d'un ensemble donné d'observations relatives à un modèle, ou ensemble de DGP. Un DGP, en tant que processus stochastique, peut être caractérisé de plusieurs manières. Nous développons maintenant la notation à partir de laquelle nous pouvons promptement exprimer une telle caractérisation qui est particulièrement utile pour nos objectifs immédiats. Nous supposons que chaque observation pour tout échantillon de taille  $n$  est une réalisation d'une variable aléatoire  $y_t$ ,  $t = 1, \dots, n$ , prenant des valeurs dans  $\mathbb{R}^m$ . Bien que la notation  $y_t$  passe sous silence la possibilité que l'observation est en général un vecteur, il est plus commode de laisser la notation vectorielle  $\mathbf{y}$  (ou  $\mathbf{y}^n$  si nous désirons faire explicitement référence à la taille de l'échantillon) désigner l'échantillon entier

$$\mathbf{y}^n = [y_1 \vdots y_2 \vdots \cdots \vdots y_n].$$

Si chaque observation est un scalaire,  $\mathbf{y}$  est un vecteur de dimension  $n$ , tandis que si chaque observation est un vecteur de dimension  $m$ ,  $\mathbf{y}$  est une matrice de dimension  $n \times m$ . Le vecteur ou la matrice  $\mathbf{y}$  peut posséder une densité de probabilité, c'est-à-dire la densité jointe de ses éléments compte tenu du DGP. Cette densité, si elle existe, est une application dont l'ensemble d'arrivée est la droite réelle et dont l'ensemble de départ est un ensemble de réalisations possibles de  $\mathbf{y}$ , ensemble que nous noterons  $\mathcal{Y}^n$  et qui sera en général un sous-ensemble de  $\mathbb{R}^{nm}$  choisi arbitrairement. Il sera nécessaire de porter toute notre attention sur la définition de la densité dans certains cas, mais il suffit pour l'instant de supposer qu'il s'agit de la densité ordinaire par rapport à la mesure de Lebesgue sur  $\mathbb{R}^{nm}$ .<sup>3</sup> Quand d'autres possibilités existent, il se trouve que le choix parmi celles-ci se révèle non pertinent pour nos propos.

Nous pouvons à présent définir formellement la fonction de vraisemblance associée à un modèle donné pour un échantillon  $\mathbf{y}$  donné. Cette fonction dépend d'une part des paramètres du modèle et d'autre part, de l'ensemble

<sup>3</sup> De cette manière, nous avons exclu les modèles à variables dépendantes qualitatives et les modèles dans lesquels la distribution de la variable dépendante a des atomes, car dans ces cas une densité par rapport à la mesure de Lebesgue n'existe pas. Voir le Chapitre 15.

d'observations donné par  $\mathbf{y}$ ; sa valeur correspond exactement à la densité associée au DGP caractérisé par le vecteur paramétrique  $\boldsymbol{\theta} \in \Theta$ , évaluée au point d'échantillon  $\mathbf{y}$ . L'ensemble  $\Theta$  désigne ici l'**espace paramétrique** dans lequel  $\boldsymbol{\theta}$  prend ses valeurs; nous supposons que c'est un sous-ensemble de  $\mathbb{R}^k$ . Nous désignerons la fonction de vraisemblance par:  $L : \mathcal{Y}^n \times \Theta \rightarrow \mathbb{R}$  et sa valeur pour  $\boldsymbol{\theta}$  et  $\mathbf{y}$  par  $L(\mathbf{y}, \boldsymbol{\theta})$ . Dans bien des cas pratiques, tel que celui examiné à la section précédente, les observations  $y_t$  sont indépendantes et chaque  $y_t$  a une densité de probabilité  $L_t(y_t, \boldsymbol{\theta})$ . La fonction de vraisemblance pour ce cas spécial est alors

$$L(\mathbf{y}, \boldsymbol{\theta}) = \prod_{t=1}^n L_t(y_t, \boldsymbol{\theta}). \quad (8.08)$$

La fonction de vraisemblance (8.03) de la section précédente est évidemment un cas particulier de ce cas présent. Quand chacune des observations  $y_t$  est identiquement distribuée selon une densité  $f(y_t, \boldsymbol{\theta})$ , comme dans cet exemple,  $L_t(y_t, \boldsymbol{\theta})$  est égale à  $f(y_t, \boldsymbol{\theta})$  pour tout  $t$ .

Même lorsque la fonction de vraisemblance ne peut pas s'écrire sous la forme de (8.08), il est toujours possible (du moins en théorie) de factoriser  $L(\mathbf{y}, \boldsymbol{\theta})$  en une série de **contributions**, chacune provenant d'une seule observation. Supposons que les observations individuelles  $y_t$ ,  $t = 1, \dots, n$ , soient *ordonnées* d'une certaine manière, comme par exemple suivant un ordre chronologique dans les séries temporelles. Or, cette factorisation peut être accomplie comme suit. Nous commençons par la densité marginale ou non conditionnelle<sup>4</sup> de la première observation  $y_1$ , que l'on peut appeler  $L_1(y_1)$ , en supprimant la dépendance par rapport à  $\boldsymbol{\theta}$  pour le moment. Puis, la densité marginale des deux premières observations jointes peut être écrite comme le produit de  $L_1(y_1)$  par la densité de  $y_2$  conditionnellement à  $y_1$ , et nous la notons  $L_2(y_2 | y_1)$ . Si maintenant, nous prenons les trois premières observations ensemble, leur densité jointe est le produit de la densité non conditionnelle des deux premières prises simultanément, par la densité de la troisième conditionnellement aux deux premières, et ainsi de suite. Le résultat pour l'échantillon

<sup>4</sup> Nous utilisons le terme "non conditionnel" par commodité. Certains statisticiens considèrent *toutes* les distributions ou *toutes* les densités comme conditionnelles à une chose ou à une autre, et nous ne voulons pas dire que nous excluons ce point de vue. Les distributions, les densités, ou espérances auxquelles nous nous référons comme non conditionnelles devraient être comprises comme étant *seulement* conditionnées aux variables véritablement exogènes, c'est-à-dire, les variables pour lesquelles le DGP est assez indépendant du DGP de  $\mathbf{y}$ . Les Bayésiens peuvent souhaiter considérer les paramètres du DGP comme des variables conditionnantes, et cette conception n'est pas non plus écartée par notre traitement.

entier des observations est

$$\begin{aligned} L(\mathbf{y}) &= L_1(y_1)L_2(y_2 | y_1)L_3(y_3 | y_2, y_1) \cdots L_n(y_n | y_{n-1}, \dots, y_1) \\ &= \prod_{t=1}^n L_t(y_t | y_{t-1}, \dots, y_1). \end{aligned} \quad (8.09)$$

Notons que ce résultat est parfaitement général et peut être appliqué à n'importe quelle densité ou fonction de vraisemblance. L'ordre des observations est habituellement l'ordre naturel, comme pour les séries temporelles, mais même si aucun ordre naturel n'existe, (8.09) demeure vraie pour un classement arbitraire.

Comme nous l'indiquions dans la dernière section, on utilise dans la pratique la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$  plutôt que la fonction de vraisemblance  $L(\mathbf{y}, \boldsymbol{\theta})$ . La décomposition de  $\ell(\mathbf{y}, \boldsymbol{\theta})$  en contributions provenant d'observations individuelles résulte de (8.09). Elle peut être écrite comme suit, en supprimant la dépendance par rapport à  $\boldsymbol{\theta}$  pour alléger les notations:

$$\ell(\mathbf{y}) = \sum_{t=1}^n \ell_t(y_t | y_{t-1}, \dots, y_1), \quad (8.10)$$

où  $\ell_t(y_t | y_{t-1}, \dots, y_1) \equiv \log L_t(y_t | y_{t-1}, \dots, y_1)$ .

Nous sommes à présent en position de donner la définition de l'**estimation par maximum de vraisemblance**. Nous disons que  $\hat{\boldsymbol{\theta}} \in \Theta$  est une estimation par maximum de vraisemblance, une **estimation ML**, ou une **MLE**, pour les données  $\mathbf{y}$  si

$$\ell(\mathbf{y}, \hat{\boldsymbol{\theta}}) \geq \ell(\mathbf{y}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta. \quad (8.11)$$

Si l'inégalité est stricte, alors  $\hat{\boldsymbol{\theta}}$  est l'unique MLE. Une MLE peut ne pas exister en général, à moins que la fonction de logvraisemblance  $\ell$  ne soit continue par rapport aux paramètres  $\boldsymbol{\theta}$ , et que l'ensemble  $\Theta$  ne soit *compact* (c'est-à-dire fermé et borné). C'est pourquoi il est d'usage, dans les traitements formels de l'estimation par maximum de vraisemblance, de supposer que  $\Theta$  est en effet compact. Nous ne désirons pas formuler cette hypothèse, parce qu'elle s'accorde en effet très mal avec la pratique standard, pour laquelle une estimation est valable partout dans  $\mathbb{R}^k$ . Mais cela signifie que nous devons vivre avec la possible non existence de la MLE.

Il est souvent commode d'utiliser une autre définition de la MLE, qui n'est pas équivalente en général. Si la fonction de vraisemblance atteint un maximum *intérieur* à l'espace paramétrique, alors elle, ou de façon équivalente la fonction de logvraisemblance, doit satisfaire les conditions du premier ordre pour un maximum. Ainsi une MLE peut se *définir* comme une solution aux **équations de vraisemblance**, qui correspondent précisément aux conditions du premier ordre suivantes:

$$\mathbf{g}(\mathbf{y}, \hat{\boldsymbol{\theta}}) \equiv \mathbf{0}, \quad (8.12)$$

où le **vecteur gradient**, ou **vecteur score**,  $\mathbf{g} \in \mathbb{R}^k$  est défini par

$$\mathbf{g}^\top(\mathbf{y}, \boldsymbol{\theta}) \equiv D_{\boldsymbol{\theta}} \ell(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^n D_{\boldsymbol{\theta}} \ell_t(\mathbf{y}, \boldsymbol{\theta}). \quad (8.13)$$

Puisque  $D_{\boldsymbol{\theta}} \ell$  est un vecteur ligne,  $\mathbf{g}$  est le vecteur *colonne* des dérivées partielles de la fonction de logvraisemblance  $\ell$  par rapport aux paramètres  $\boldsymbol{\theta}$ . Nous avons écrit  $\ell_t(\mathbf{y}, \boldsymbol{\theta})$ , et non  $\ell_t(y_t, \boldsymbol{\theta})$ , parce qu'en général  $\ell_t$  peut dépendre de valeurs "passées" de la variable dépendante,  $y_{t-1}, y_{t-2}, \dots$ . Elle ne dépend pas des valeurs "futures" bien entendu, mais l'utilisation de la notation vectorielle est encore le moyen le plus simple de nous rappeler de la dépendance par rapport à d'autres éléments que  $y_t$ .

Comme il peut arriver que plus d'une valeur de  $\boldsymbol{\theta}$  satisfasse les équations de vraisemblance (8.12), la définition nécessite par ailleurs que l'estimation  $\hat{\boldsymbol{\theta}}$  soit associée à un *maximum* local de  $\ell$  et que

$$\text{plim}_{n \rightarrow \infty} (n^{-1} \ell(\mathbf{y}, \hat{\boldsymbol{\theta}})) \geq \text{plim}_{n \rightarrow \infty} (n^{-1} \ell(\mathbf{y}, \boldsymbol{\theta}^*)),$$

où  $\boldsymbol{\theta}^*$  est n'importe quelle autre solution des équations de vraisemblance. Cette seconde définition de la MLE est souvent associée à Cramér, dans sa célèbre preuve de convergence (Cramér, 1946). Dans la pratique, la nécessité que  $\text{plim}(n^{-1} \ell(\mathbf{y}, \hat{\boldsymbol{\theta}})) \geq \text{plim}(n^{-1} \ell(\mathbf{y}, \boldsymbol{\theta}^*))$  est à l'évidence impossible à vérifier en général. Le problème vient du fait que l'on ne connaît pas le DGP et que par conséquent, le calcul analytique des limites en probabilité est impossible. Si pour un échantillon donné il existe deux racines ou plus aux équations de vraisemblance, celle qui est associée à la valeur la plus haute de  $\ell(\mathbf{y}, \boldsymbol{\theta})$  pour cet échantillon peut ne pas converger vers celle qui est associée à la valeur la plus haute asymptotiquement. Dans la pratique, s'il existe plus d'une solution pour les équations de vraisemblance, l'on sélectionne celle qui est associée à la valeur la plus haute de la fonction de logvraisemblance. Malgré tout, s'il y a deux ou plusieurs solutions pour lesquelles les valeurs correspondantes de  $\ell(\mathbf{y}, \boldsymbol{\theta})$  sont très proches, il est fort possible de sélectionner la mauvaise.

Nous insistons sur le fait que ces deux définitions de la MLE ne sont pas équivalentes. En conséquence, il est parfois nécessaire de parler des **MLE du Type 1** quand nous faisons référence à celles obtenues par la maximisation de  $\ell(\mathbf{y}, \boldsymbol{\theta})$  sur  $\Theta$ , et des **MLE de Type 2** quand nous faisons référence à celles obtenues comme solutions des équations de vraisemblance. Bien que dans la plupart des cas, en pratique, chacune pourrait être utilisée et que dans certains cas, les deux types de MLE coïncident, il existe des situations où seul un des deux types de MLE est réalisable. En particulier, il existe des modèles où  $\ell(\boldsymbol{\theta})$  est non bornée dans certaines directions, et la définition de l'estimateur de Type 1 ne peut donc pas être utilisée, mais néanmoins il existe un  $\hat{\boldsymbol{\theta}}$  qui est une racine convergente des équations de vraisemblance; consulter

Kiefer (1978) pour un modèle de ce genre. D'un autre côté, la définition de l'estimateur de Type 2 ne s'applique pas au problème standard de l'estimation d'un ou de deux points terminaux d'une distribution uniforme, parce que les équations de vraisemblance ne sont jamais satisfaites.

Il est utile d'étudier le problème de l'estimation des points terminaux d'une distribution uniforme. Supposons que pour tout  $t$  la densité de  $y_t$  soit

$$f(y_t) = \begin{cases} 1/\alpha & \text{si } 0 \leq y_t \leq \alpha \\ 0 & \text{sinon.} \end{cases}$$

Ici, on sait qu'une borne de la distribution uniforme est zéro, mais il faut estimer  $\alpha$ , l'autre borne. Les fonctions de vraisemblance et de logvraisemblance sont respectivement,

$$L(\mathbf{y}, \alpha) = \begin{cases} \alpha^{-n} & \text{si } 0 \leq y_t \leq \alpha \text{ pour tout } y_t \\ 0 & \text{sinon} \end{cases}$$

et

$$\ell(\mathbf{y}, \alpha) = \begin{cases} -n \log(\alpha) & \text{si } 0 \leq y_t \leq \alpha \text{ pour tout } y_t \\ -\infty & \text{sinon.} \end{cases} \quad (8.14)$$

L'équation de vraisemblance obtenue en dérivant  $\ell(\mathbf{y}, \alpha)$  par rapport à  $\alpha$  et en annulant la dérivée est

$$-\frac{n}{\alpha} = 0.$$

Comme cette équation n'a pas de solution finie, il n'existe aucune estimation ML de Type 2. Cependant, il est clair que nous pouvons trouver une estimation ML de Type 1. De (8.14), il est évident que pour maximiser  $\ell(\mathbf{y}, \alpha)$  nous devons rendre  $\hat{\alpha}$  aussi petite que possible. Comme  $\hat{\alpha}$  ne peut pas être plus petite que la plus grande valeur de  $y_t$  observée, l'estimation ML de Type 1 doit simplement être

$$\hat{\alpha} = \max_t(y_t).$$

Par le terme **estimateur du maximum de vraisemblance** nous désignerons la variable aléatoire qui associe à chaque occurrence aléatoire possible  $\mathbf{y}$  la MLE correspondante.<sup>5</sup> La distinction entre une **estimation** et un **estimateur** a été établie dans la Section 5.2. Nous pouvons rappeler qu'un estimateur, une variable aléatoire, est représenté comme une fonction (implicite ou explicite) des ensembles possibles d'observations, alors qu'une estimation est une valeur que peut prendre cette fonction pour un ensemble d'observations bien spécifié.

<sup>5</sup> Dans les cas de non existence de la MLE dans certains échantillons, l'estimateur peut être défini comme une variable aléatoire appropriée en lui assignant une valeur arbitrairement, telle que  $-\infty$ , pour ces échantillons où la MLE n'existe pas.



Tout comme il existe deux définitions possibles des estimations ML, il existe également deux définitions possibles d'un estimateur ML. Les définitions suivantes montrent clairement que l'estimateur est une variable aléatoire, qui dépend des valeurs observées de l'échantillon  $\mathbf{y}$ . L'**estimateur de Type 1**, correspondant à la définition standard (8.11) de la MLE, est  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  défini par

$$L(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) > L(\mathbf{y}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \text{ tel que } \boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}(\mathbf{y}). \quad (8.15)$$

L'**estimateur de Type 2**, correspondant à la définition (8.12) de Cramér, est  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  défini par:

$$\mathbf{g}(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) = \mathbf{0}, \quad (8.16)$$

où  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  donne un maximum local de  $\ell$ , et

$$\text{plim}_{n \rightarrow \infty} \left( n^{-1} \ell(\mathbf{y}, \hat{\boldsymbol{\theta}}(\mathbf{y})) \right) \geq \text{plim}_{n \rightarrow \infty} \left( n^{-1} \ell(\mathbf{y}, \boldsymbol{\theta}^*(\mathbf{y})) \right) \quad (8.17)$$

pour n'importe quelle autre solution  $\boldsymbol{\theta}^*(\mathbf{y})$  des équations de vraisemblance.

Nous concluons cette section par une variété de définitions qui seront utilisées dans le reste du chapitre et plus généralement dans le reste du livre. En utilisant la décomposition (8.10) de la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$ , nous pouvons définir une matrice  $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$  de dimension  $n \times k$  dont l'élément type est

$$G_{ti}(\mathbf{y}, \boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_i}. \quad (8.18)$$

Nous appellerons  $\mathbf{G}(\mathbf{y}, \boldsymbol{\theta})$  la **matrice des contributions au gradient**, ou **matrice CG** pour faire court. Cette matrice est intimement reliée au vecteur gradient  $\mathbf{g}$ , qui est juste  $\mathbf{G}^T \boldsymbol{\iota}$ , où comme d'habitude  $\boldsymbol{\iota}$  désigne un vecteur de taille  $n$  pour lequel chaque élément est égal à 1. La  $t^{\text{ième}}$  ligne de  $\mathbf{G}$ , qui mesure la contribution au gradient de la  $t^{\text{ième}}$  observation, sera noté  $\mathbf{G}_t$ .

La **matrice Hessienne** associée à la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$  est la matrice  $\mathbf{H}(\mathbf{y}, \boldsymbol{\theta})$  de dimension  $k \times k$  dont l'élément type est

$$H_{ij}(\mathbf{y}, \boldsymbol{\theta}) \equiv \frac{\partial^2 \ell(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}. \quad (8.19)$$

Nous définissons l'**espérance de la Hessienne moyenne** pour un échantillon de taille  $n$  comme

$$\mathcal{H}^n(\boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}}(n^{-1} \mathbf{H}(\mathbf{y}, \boldsymbol{\theta})).$$

La notation  $E_{\boldsymbol{\theta}}$  signifie que l'espérance est calculée en utilisant le DGP caractérisé par le vecteur paramétrique  $\boldsymbol{\theta}$  plutôt que par le DGP qui pourrait réellement avoir généré un quelconque échantillon particulier donné. Ainsi, un DGP différent est implicitement utilisé pour calculer l'espérance pour chaque

$\theta$ . La **limite de la Hessienne** ou **Hessienne asymptotique**, si elle existe, est définie comme

$$\mathcal{H}(\theta) \equiv \lim_{n \rightarrow \infty} \mathcal{H}^n(\theta).$$

Cette quantité, qui est une matrice symétrique, et en général semi-définie négative, apparaîtra un grand nombre de fois dans la théorie asymptotique de l'estimation ML.

Nous définissons l'**information contenue dans l'observation  $t$**  par  $\mathbf{I}_t(\theta)$ , la matrice de dimension  $k \times k$  dont l'élément type est

$$(\mathbf{I}_t(\theta))_{ij} \equiv E_\theta(G_{ti}(\theta)G_{tj}(\theta)). \quad (8.20)$$

Le fait que  $\mathbf{I}_t(\theta)$  soit une matrice symétrique, en général semi-définie positive, et qu'elle soit définie positive à condition qu'il existe une relation linéaire entre les composantes du vecteur aléatoire  $\mathbf{G}_t$  est une conséquence immédiate de cette définition. La **matrice d'information moyenne** pour un échantillon de taille  $n$  est définie par

$$\mathcal{J}^n(\theta) \equiv \frac{1}{n} \sum_{t=1}^n \mathbf{I}_t(\theta) = n^{-1} \mathbf{I}^n, \quad (8.21)$$

et la **matrice d'information à la limite** ou **matrice d'information asymptotique**, si elle existe, est définie par

$$\mathcal{J}(\theta) \equiv \lim_{n \rightarrow \infty} \mathcal{J}^n(\theta). \quad (8.22)$$

La matrice  $\mathbf{I}_t(\theta)$  mesure la quantité *espérée* d'information contenue dans la  $t^{\text{ième}}$  observation et  $\mathbf{I}^n \equiv n\mathcal{J}^n$  mesure la quantité espérée d'information contenue dans l'échantillon entier. Les matrices d'information  $\mathcal{J}^n$  et  $\mathcal{J}$  sont, comme  $\mathbf{I}_t$ , symétriques, et en général semi-définies positives. La matrice d'information moyenne  $\mathcal{J}^n$  et l'espérance de la Hessienne moyenne  $\mathcal{H}^n$  ont été définies telles qu'elles soient  $O(1)$  quand  $n \rightarrow \infty$ . Elles sont donc très pratiques à utiliser lors de l'analyse asymptotique. La terminologie dans ce domaine n'est pas entièrement unifiée. Certains auteurs utilisent simplement le terme "matrice d'information" pour se référer à  $\mathcal{J}^n$ , tandis que d'autres l'utilisent pour se référer à  $n$  fois  $\mathcal{J}^n$ , ce que nous avons appelé  $\mathbf{I}^n$ .

### 8.3 TRANSFORMATIONS ET REPARAMÉTRISATIONS

Dans cette section et dans les suivantes, nous développons la théorie classique de l'estimation par maximum de vraisemblance et, en particulier, nous démontrons les propriétés qui font que cette théorie produit une méthode d'estimation qui possède de nombreux avantages. Nous démontrerons aussi que dans certaines circonstances ces propriétés font défaut. Comme nous en

avons discuté dans la Section 8.1, les principales caractéristiques enviables des estimateurs ML sont l'**invariance**, la **convergence**, la **normalité asymptotique**, l'**efficacité asymptotique**, et la **calculabilité**. Dans cette section, nous discuterons de la première de celles-ci, l'invariance des estimateurs ML à la reparamétrisation du modèle.

L'idée d'invariance est un concept important dans l'analyse économétrique. Notons  $\mathbb{M}$  le modèle qui nous intéresse. Une **paramétrisation** du modèle  $\mathbb{M}$  est une application, disons  $\lambda$ , dont l'espace de départ est un espace paramétrique  $\Theta$  et qui va vers  $\mathbb{M}$ . Il existera en général une infinité de paramétrisations pour tout modèle  $\mathbb{M}$  donné. Après tout, peu de contraintes portent sur l'espace paramétrique  $\Theta$ , en dehors de sa dimension. Il est possible de construire une application bijective et dérivable partant d'un sous-ensemble de  $\mathbb{R}^k$  vers pratiquement n'importe quel autre sous-ensemble de  $\mathbb{R}^k$  par des procédés tels que la translation, la rotation, la dilatation, et bien d'autres encore, et n'importe lequel de ces autres sous-ensembles peut donc faire office d'espace paramétrique pour le modèle  $\mathbb{M}$ . C'est justement à cause de ces possibilités, que l'on désire que les estimateurs possèdent la propriété d'invariance. Le terme d'"invariance" est compris dans ce contexte comme l'invariance au type de transformation dont nous avons discuté, et que nous appelons formellement **reparamétrisation**.

Pour illustrer le fait que n'importe quel modèle peut être paramétrisé un nombre infini de fois, considérons le cas d'une distribution exponentielle, dont nous avons discuté dans la Section 8.1. Nous avons vu que la fonction de vraisemblance pour un échantillon de réalisations indépendantes obéissant à cette distribution était (8.03). Si nous posons  $\theta \equiv \delta^\alpha$ , nous pouvons définir une famille entière de paramétrisations indexées par  $\alpha$ . Nous pouvons choisir  $\alpha$  comme étant n'importe quel nombre fini non nul. La fonction de vraisemblance correspondant à cette famille de paramétrisations est

$$L(\mathbf{y}, \delta) = \prod_{t=1}^n \delta^\alpha e^{-\delta^\alpha y_t}.$$

Evidemment, le cas  $\alpha = 1$  correspond à la paramétrisation en  $\theta$  de (8.02) et le cas  $\alpha = -1$  correspond à la paramétrisation en  $\phi$  de (8.07).

Il est facile de voir que les estimateurs ML sont invariants aux reparamétrisations du modèle. Définissons par  $\boldsymbol{\eta} : \Theta \rightarrow \Phi \subseteq \mathbb{R}^k$  une application régulière qui transforme le vecteur  $\boldsymbol{\theta}$  en un unique vecteur  $\boldsymbol{\phi} \equiv \boldsymbol{\eta}(\boldsymbol{\theta})$ . La fonction de vraisemblance pour le modèle  $\mathbb{M}$  en termes des nouveaux paramètres  $\boldsymbol{\phi}$ , disons  $L'$ , est définie par la relation

$$L'(\mathbf{y}, \boldsymbol{\phi}) = L(\mathbf{y}, \boldsymbol{\theta}) \quad \text{où } \boldsymbol{\phi} = \boldsymbol{\eta}(\boldsymbol{\theta}). \quad (8.23)$$

L'équation (8.23) suit immédiatement des faits que la fonction de vraisemblance est la densité d'un processus stochastique et que  $\boldsymbol{\theta}$  et  $\boldsymbol{\phi} = \boldsymbol{\eta}(\boldsymbol{\theta})$



parce que, pour une fonction non linéaire  $\boldsymbol{\eta}(\boldsymbol{\theta})$ ,

$$E_0(\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})) \neq \boldsymbol{\eta}(E_0(\hat{\boldsymbol{\theta}})) = \boldsymbol{\eta}(\boldsymbol{\theta}_0) = \boldsymbol{\phi}_0.$$

Ceci suggère que, bien que la paramétrisation que nous choisissons n'ait pas d'importance pour l'estimation du DGP, elle peut avoir un effet substantiel sur les propriétés de nos estimations paramétriques avec des échantillons finis. En choisissant la paramétrisation appropriée, nous pouvons dans certains cas assurer que nos estimations sont sans biais, ou proches d'être sans biais, et que leurs distributions sont proches de leurs distributions asymptotiques. Par contraste, si nous choisissons une paramétrisation inappropriée, nous pourrions par inadvertance rendre nos estimations sévèrement biaisées et dont les distributions sont éloignées de leurs distributions asymptotiques.

## 8.4 LA CONVERGENCE

Une des raisons pour lesquelles l'estimation par maximum de vraisemblance est largement utilisée est que les estimateurs ML sont, sous des conditions assez générales, convergents. Dans cette section, nous expliquons pourquoi c'est le cas. Nous nous intéressons premièrement à l'estimateur ML de Type 1, bien que nous proposons aussi certaines discussions au sujet de l'estimateur de Type 2. Nous commençons en posant la définition:

$$\bar{\ell}(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \equiv \text{plim}_0(n^{-1}\ell^n(\mathbf{y}^n, \boldsymbol{\theta})), \quad (8.24)$$

où la notation "plim<sub>0</sub>" signifie comme d'habitude que la limite en probabilité est calculée sous le DGP caractérisé par  $\boldsymbol{\theta}_0$ . La fonction  $\bar{\ell}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$  est la valeur limite de  $n^{-1}$  fois la fonction de logvraisemblance, quand les données sont générées par un cas particulier du modèle avec  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . Une condition de régularité importante qui doit être satisfaite afin qu'un estimateur ML soit convergent est que le modèle soit asymptotiquement identifié. Par définition, ceci sera le cas si le problème

$$\max_{\boldsymbol{\theta} \in \Theta} \bar{\ell}(\boldsymbol{\theta}; \boldsymbol{\theta}_0) \quad (8.25)$$

ne comporte qu'une unique solution. Cette définition implique que n'importe quel DGP appartenant au modèle générera des échantillons qui, s'ils sont suffisamment grands, identifieront le modèle. L'interprétation est la même que dans le contexte du modèle de régression.

Nous désirons maintenant démontrer que  $\bar{\ell}(\boldsymbol{\theta}; \boldsymbol{\theta}_0)$  est maximale en  $\boldsymbol{\theta}_0$ , la valeur de  $\boldsymbol{\theta}$  qui caractérise le DGP. Nous désignons par  $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\mathbf{y})$  le maximum global de la fonction de vraisemblance  $L(\mathbf{y}, \boldsymbol{\theta})$ , et réclamons que cette fonction soit *continue* en  $\boldsymbol{\theta}$ , et nous désignons par  $\boldsymbol{\theta}^*$  n'importe quel autre vecteur de paramètres (non stochastique) dans  $\Theta$ , et réclamons que cet espace soit

compact. Ces deux exigences signifient qu'il n'y a aucun problème sur la possible non existence de la MLE. Nous désignerons les espérances calculées par rapport au DGP par  $E_0(\cdot)$ . Alors, grâce à l'inégalité de Jensen (consulter l'Annexe B), on montre que

$$E_0\left(\log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right)\right) \leq \log\left(E_0\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right)\right), \quad (8.26)$$

car le logarithme est une fonction concave. Plus loin, (8.26) deviendra une inégalité stricte à chaque fois que  $L(\boldsymbol{\theta}^*)/L(\boldsymbol{\theta}_0)$  sera une variable aléatoire non dégénérée. Une dégénérescence se produira seulement s'il existe  $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$  tel que  $L(\boldsymbol{\theta}')/L(\boldsymbol{\theta}_0)$  soit identiquement unitaire;  $\ell(\boldsymbol{\theta}') - \ell(\boldsymbol{\theta}_0)$  serait alors identiquement égale à zéro. Mais la condition d'identification asymptotique (8.25) élimine cette possibilité pour des tailles d'échantillon assez grandes, puisque, si elle est vérifiée,  $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_0$  implique que  $L(\boldsymbol{\theta}') \neq L(\boldsymbol{\theta}_0)$ .

En utilisant le fait que  $L(\boldsymbol{\theta}_0)$  est la densité jointe de  $\mathbf{y}$ , nous voyons que l'espérance à l'intérieur du logarithme dans le membre de droite de (8.26) est

$$E_0\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right) = \int_{\mathbf{y}^n} \frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)} L(\boldsymbol{\theta}_0) d\mathbf{y} = \int_{\mathbf{y}^n} L(\boldsymbol{\theta}^*) d\mathbf{y} = 1.$$

Nous gérerons la nullité éventuelle de  $L(\boldsymbol{\theta}_0)$  en définissant la seconde intégrale ci-dessus comme nulle lorsque  $L(\boldsymbol{\theta}_0)$  l'est aussi. Comme le logarithme de 1 est 0, il suit de (8.26) que

$$E_0\left(\log\left(\frac{L(\boldsymbol{\theta}^*)}{L(\boldsymbol{\theta}_0)}\right)\right) < 0,$$

qui peut être récrit comme

$$E_0(\ell(\boldsymbol{\theta}^*)) - E_0(\ell(\boldsymbol{\theta}_0)) < 0. \quad (8.27)$$

Ainsi, l'espérance de la fonction de logvraisemblance lorsqu'elle est évaluée avec le véritable vecteur paramétrique,  $\boldsymbol{\theta}_0$ , est strictement supérieure à l'espérance évaluée avec n'importe quel autre vecteur de paramètres,  $\boldsymbol{\theta}^*$ .

La prochaine étape consiste à montrer que ce qui est vrai pour les espérances mathématiques dans (8.27), l'est aussi, à la limite lorsque  $n \rightarrow \infty$ , pour l'analogie correspondant à l'échantillon. Cette expression analogue correspondant à l'échantillon est

$$\frac{1}{n}(\ell(\boldsymbol{\theta}^*) - \ell(\boldsymbol{\theta}_0)) = \frac{1}{n} \sum_{t=1}^n \ell_t(\mathbf{y}, \boldsymbol{\theta}^*) - \frac{1}{n} \sum_{t=1}^n \ell_t(\mathbf{y}, \boldsymbol{\theta}_0). \quad (8.28)$$

Maintenant, il est nécessaire de supposer que les sommes dans (8.28) satisfont certaines conditions de régularité suffisantes pour qu'une loi des grands

nombres leur soit appliquée. Comme nous l'avons vu dans le Chapitre 4, celles-ci nécessitent que les  $\ell_t$  soient indépendantes ou du moins, qu'elles ne manifestent pas trop fortement une dépendance; qu'elles possèdent une sorte d'espérance (bien qu'elles puissent ne pas posséder une espérance habituelle); et qu'elles possèdent des variances bornées supérieurement; pour tous les détails, consulter la Section 4.7. Nous pouvons donc réclamer, parce que cela est pratique, que pour tout  $\theta \in \Theta$ ,  $\{\ell_t(\theta)\}_{t=1}^\infty$  satisfait la condition WULLN de la Section 4.7 pour le DGP caractérisé par  $\theta_0$ . Nous pouvons alors utiliser (8.27) pour affirmer que

$$\text{plim}_0(n^{-1}\ell(\theta^*)) - \text{plim}_0(n^{-1}\ell(\theta_0)) < 0, \quad (8.29)$$

où les deux limites en probabilité existent. En fait, grâce à la définition (8.24),

$$\text{plim}_0(n^{-1}\ell(\theta^*)) = \bar{\ell}(\theta^*; \theta_0),$$

ce qui démontre l'existence de la fonction  $\bar{\ell}(\theta^*; \theta_0)$ . Il reste à démontrer que l'inégalité dans (8.29) est stricte, car la *limite* des inégalités strictes (8.27) n'est pas nécessairement une inégalité stricte. Cependant, la condition d'identification asymptotique (8.25) peut encore être invoquée pour rétablir l'inégalité stricte.

Avec l'hypothèse d'identification asymptotique donnée et le résultat (8.29), il est maintenant facile de voir pourquoi  $\hat{\theta}$  doit être convergente. Nous savons que

$$n^{-1}\ell(\hat{\theta}) \geq n^{-1}\ell(\theta_0), \quad (8.30)$$

pour tout  $n$ , parce que  $\hat{\theta}$  maximise la fonction de logvraisemblance. Clairement (8.29) et (8.30) ne peuvent pas toutes deux être vraies à moins que

$$\text{plim}_0(n^{-1}\ell(\hat{\theta})) = \text{plim}_0(n^{-1}\ell(\theta_0)). \quad (8.31)$$

Mais si le modèle est asymptotiquement identifié, la valeur  $\hat{\theta}$  qui maximise (8.24) doit être unique. Alors, (8.31) ne peut pas être vérifiée à moins que  $\text{plim}_0(\hat{\theta}) = \theta_0$ .<sup>6</sup>

Nous pouvons maintenant énoncer le théorème suivant, que l'on doit à Wald (1949):

*Théorème 8.1. Théorème de Convergence de Wald.*

L'estimateur ML (8.15) pour un modèle représenté par la famille paramétrique des fonctions de logvraisemblance  $\ell(\theta)$  dans lesquelles  $\theta$  est contraint à résider dans un espace paramétrique compact, est convergent si les contributions  $\{\ell_t(\theta)\}_{t=1}^\infty$  satisfont les conditions de

<sup>6</sup> Parce que  $\hat{\theta}$  est stochastique, cet argument n'est pas rigoureux.

régularité WULLN et si, en plus, le modèle est asymptotiquement identifié.

Notons que le résultat a été démontré uniquement pour des espaces paramétriques *compacts*, car autrement nous ne pourrions pas être sûr que  $\hat{\theta}$  existe pour tout  $n$ . Il existe des modèles, par exemple certains appelés modèles de régime endogène, dans lesquels le fait qu'une variance ne puisse tendre vers zéro pour une densité de probabilité qui a de bonnes propriétés, conduit à une défaillance de la compacité de l'espace paramétrique (puisque'en excluant une variance nulle, on crée une borne ouverte partiellement dans cet espace). Par exemple, il peut ne pas exister de MLE de Type 1 avec une limite en probabilité; consulter Kiefer (1978).

Il existe deux ensembles majeurs de circonstances dans lesquelles les estimations ML peuvent ne pas être convergentes. Le premier survient quand le nombre de paramètres n'est pas fixe mais augmente avec  $n$ . Cette possibilité n'est même pas considérée dans le théorème précédent, où  $\theta$  est indépendant de  $n$ . Mais il n'est pas surprenant que cela engendre des problèmes, car si le nombre de paramètres n'est pas fixe, il est loin d'être évident que la quantité d'information que l'échantillon nous donne à propos de chacun d'eux augmentera suffisamment rapidement lorsque  $n \rightarrow \infty$ . Il est en fait possible de laisser le nombre de paramètres augmenter, mais le taux d'accroissement doit être modéré (par exemple, comme  $n^{1/4}$ ). De tels problèmes sont bien au-delà des objectifs de cet ouvrage; consulter, entre d'autres, Neyman et Scott (1948), Kiefer et Wolfowitz (1956), et Kalbfleisch et Sprott (1970).

Les cas d'absence de convergence les plus fréquemment rencontrés sont ceux dans lesquels le modèle n'est pas identifié asymptotiquement. Ceci peut arriver même quand il *est* identifié par n'importe quel échantillon fini. Par exemple, considérons le modèle de régression

$$y_t = \alpha \frac{1}{t} + u_t, \quad u_t \sim \text{NID}(0, 1),$$

considéré à l'origine dans la Section 5.2. Nous avons déjà vu que des modèles de ce type ne peuvent pas être estimés de manière convergente par les moindres carrés, et c'est un exercice simple de montrer que de tels modèles ne peuvent pas non plus être estimés de manière convergente par le maximum de vraisemblance. Une manière de concevoir ce type de problème est d'observer que, lorsque  $n$  augmente, chaque observation nouvelle porte de moins en moins d'information au sujet de  $\alpha$ . Ainsi, bien que la matrice d'information d'échantillon fini  $\mathbf{I}^n$  soit toujours de plein rang (de un dans ce cas), la matrice d'information asymptotique  $\mathcal{I}$  ne l'est pas (elle converge vers zéro dans ce cas). Dans ce cas habituel où l'estimateur ML est convergent, chaque nouvelle observation additionne approximativement la même quantité d'information et  $\mathcal{I}$ , étant la limite de la moyenne des  $\mathbf{I}_t$ , sera alors de plein rang.

Dans la plupart des situations, la seule chose que nous aurons besoin de connaître sera la convergence de l'estimateur ML de Type 1. Cependant, on



trouve des cas dans lesquels seul l'estimateur de Type 2 existe. Dans le reste de cette section, nous esquissons alors la preuve de la convergence de l'estimateur ML de Type 2, tel qu'il est défini par (8.16) et (8.17). Pour que cet estimateur existe, il est bien sûr nécessaire que les contributions  $\ell_t$  pour la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$  soient dérivables par rapport aux paramètres  $\boldsymbol{\theta}$ , et aussi supposons-nous qu'elles sont continûment différentiables au moins une fois. Grâce à cette hypothèse, l'argument qui suit n'est plus utile dans de nombreux ensembles de circonstances: si l'espace paramétrique  $\Theta$  est compact et le vecteur paramétrique  $\boldsymbol{\theta}_0$  associé au DGP est à l'intérieur de  $\Theta$ , alors pour des échantillons assez importants, la probabilité que la maximum de  $\ell$  soit réalisé en un point intérieur de  $\Theta$  devient arbitrairement proche de l'unité. Quand cela arrive, les estimateurs de Type 1 et de Type 2 coïncideront asymptotiquement. D'un autre côté, si  $\boldsymbol{\theta}_0$  est sur la frontière de  $\Theta$ , il y aura une probabilité positive, pour des échantillons arbitrairement grands, que l'estimateur de Type 2 n'existe pas. Dans un tel cas, la question de sa convergence éventuelle ne se pose pas.

La situation est plus délicate dans le cas d'un espace paramétrique non compact. Nous remarquons tout d'abord que si  $\boldsymbol{\theta}_0$  se situe sur la frontière de  $\Theta$ , il y aura une probabilité positive pour que l'estimateur de Type 2 n'existe pas, mais ce n'est pas la compacité qui est en cause. Nous supposons donc que  $\boldsymbol{\theta}_0$  est à l'intérieur de  $\Theta$ . Nous supposons ensuite que la condition de la définition suivante est satisfaite:

*Définition 8.1.*

Le modèle caractérisé par la fonction de logvraisemblance  $\ell$  est **identifiée asymptotiquement sur un espace paramétrique  $\Theta$  non compact** si le modèle est asymptotiquement identifié et si, de plus, il n'existe aucune séquence  $\{\boldsymbol{\theta}^n\}$  ne comportant aucun point limite qui satisfasse

$$\bar{\ell}(\boldsymbol{\theta}^n; \boldsymbol{\theta}_0) \longrightarrow \bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0); \quad \bar{\ell}(\boldsymbol{\theta}^n; \boldsymbol{\theta}_0) < \bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0). \quad (8.32)$$

L'identification asymptotique semble écarter l'existence de telles séquences, mais il n'en est rien. Pour que la séquence n'ait aucun point limite, elle doit diverger à l'infini dans certaines directions, ou autrement, converger vers un point qui n'appartient pas à l'espace paramétrique non compact, tel qu'un point de variance nulle. Ainsi, le fait que  $\bar{\ell}(\boldsymbol{\theta}^n; \boldsymbol{\theta}_0)$  tende vers la limite  $\bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$  n'implique pas l'existence d'un point dans  $\Theta$ , disons  $\boldsymbol{\theta}^\infty$ , pour lequel  $\bar{\ell}(\boldsymbol{\theta}^\infty; \boldsymbol{\theta}_0) = \bar{\ell}(\boldsymbol{\theta}_0; \boldsymbol{\theta}_0)$ . En effet, l'existence de  $\boldsymbol{\theta}^\infty$  devrait contredire l'identification asymptotique dans son sens habituel. Mais pour que l'on puisse interpréter l'identification asymptotique dans son sens habituel dans un espace paramétrique non compact, l'existence de suites satisfaisant (8.32) doit être éliminée, même si elles n'ont pas de point limite.

Retournons maintenant au cas des estimateurs de Type 2. Considérons un voisinage *compact*  $\Theta_0$  de  $\boldsymbol{\theta}_0$ . Nous pourrions définir un autre estimateur ML comme le point qui donne le maximum de  $\ell$  dans  $\Theta_0$ . Grâce au

Théorème de convergence de Wald (Théorème 8.1) ce nouvel estimateur serait convergent. Deux cas possibles semblent alors exister. Le premier est celui pour lequel il existe une probabilité positive asymptotiquement que cet estimateur soit sur la *frontière* du voisinage  $\Theta_0$  et le second est celui pour lequel cette probabilité est nulle. Dans le second cas, le nouvel estimateur et l'estimateur de Type 2 coïncident asymptotiquement, compte tenu de la condition d'identification asymptotique pour un ensemble non compact  $\Theta$ , et ce dernier est donc convergent. Mais en fait le premier cas ne peut pas survenir. Pour un  $\Theta_0$  fixé,  $\theta_0$  est à une distance positive de la frontière de  $\Theta_0$ , et la convergence du nouvel estimateur exclut toute probabilité positive asymptotiquement concentrée sur une région fermée éloignée de  $\theta_0$ . Ainsi nous concluons que lorsque l'espace paramétrique est non compact, à condition que le DGP reste à l'intérieur de cet espace et que le modèle soit asymptotiquement identifié sur son espace paramétrique non compact, l'estimateur de Type 2 est convergent. Ces résultats sont résumés dans le théorème suivant:

*Théorème 8.2. Second Théorème de Convergence.*

Soit un modèle représenté par une famille paramétrique de fonctions de logvraisemblance  $\ell(\theta)$  au moins une fois continûment différentiables dans laquelle  $\theta$  est contraint d'appartenir à un espace paramétrique non nécessairement compact. Alors, pour les DGP qui se situent à l'intérieur de cet espace paramétrique, l'estimateur ML défini par (8.16) et (8.17) est convergent si les contributions  $\{\ell_t(\theta)\}_{t=1}^\infty$  satisfont les conditions de régularité WULLN et si de plus l'espace paramétrique est compact et le modèle est asymptotiquement identifié, ou si l'espace paramétrique est non compact et le modèle est asymptotiquement identifié au sens de la Définition 8.1.

## 8.5 LA DISTRIBUTION ASYMPTOTIQUE DE L'ESTIMATEUR ML

Nous commençons notre analyse en démontrant un résultat simple mais fondamental concernant le gradient  $\mathbf{g}$  et la matrice  $\mathbf{G}$  de CG:

$$E_\theta(G_{ti}(\theta)) \equiv E_\theta\left(\frac{\partial \ell_t(\theta)}{\partial \theta_i}\right) = 0. \quad (8.33)$$

Ce résultat indique que, sous le DGP caractérisé par  $\theta$ , l'espérance de chaque élément de la matrice CG, évaluée en  $\theta$ , est zéro. Ceci implique que

$$E_\theta(\mathbf{g}(\theta)) = \mathbf{0} \quad \text{et} \quad E_\theta(\mathbf{G}(\theta)) = \mathbf{0}.$$

C'est un résultat très important pour plusieurs raisons. En particulier, il nous permettra d'appliquer un théorème de la limite centrale à la quantité

$n^{-1/2}g(\theta_0)$ . La démonstration est comme suit:

$$\begin{aligned}
E_\theta(G_{ti}(y_t, \theta)) &= \int \frac{\partial \log L_t(y_t, \theta)}{\partial \theta_i} L_t(y_t, \theta) dy_t \\
&= \int \frac{1}{L_t(y_t, \theta)} \frac{\partial L_t(y_t, \theta)}{\partial \theta_i} L_t(y_t, \theta) dy_t \\
&= \int \frac{\partial L_t(y_t, \theta)}{\partial \theta_i} dy_t \\
&= \frac{\partial}{\partial \theta_i} \int L_t(y_t, \theta) dy_t \\
&= \frac{\partial}{\partial \theta_i} (1) = 0.
\end{aligned} \tag{8.34}$$

L'avant dernière étape est simplement une conséquence de la normalisation de la densité  $L_t(y_t, \theta)$ . L'étape précédente, dans laquelle les ordres de différentiation et d'intégration sont interchangés, est valide sous une variété de conditions de régularité, parmi lesquelles la plus simple est que le domaine d'intégration, disons  $\mathcal{Y}_t$ , soit indépendant de  $\theta$ . De façon alternative, si cette hypothèse n'est pas vraie, alors il suffit que  $L_t(y_t, \theta)$  s'annule sur la frontière du domaine  $\mathcal{Y}_t$  et que  $\partial \ell_t(y_t, \theta)/\partial \theta$  soit uniformément bornée; consulter l'Annexe B.

Les résultats simples concernant la distribution asymptotique des estimations ML sont obtenus le plus facilement dans le contexte de l'estimateur de Type 2, défini par (8.16) et (8.17). Par conséquent, nous limiterons notre attention à ce cas et nous supposons que  $\hat{\theta}$  est une racine des équations de vraisemblance (8.12). Il est alors relativement simple de montrer que  $\hat{\theta}$  possède la propriété de **normalité asymptotique**, dont nous avons discuté dans le Chapitre 5. Pour un DGP caractérisé par  $\theta_0$ , le vecteur des estimations paramétriques  $\hat{\theta}$  tend vers la limite non stochastique  $\theta_0$ . Cependant, si nous multiplions la différence  $\hat{\theta} - \theta_0$  par  $n^{1/2}$ , la quantité résultante  $n^{1/2}(\hat{\theta} - \theta_0)$  aura une limite en probabilité qui est une variable aléatoire avec une distribution normale multivariée. Comme dans le cas des NLS, nous pouvons occasionnellement y faire référence de façon peu formelle comme à la distribution asymptotique de  $\hat{\theta}$ , bien que cela ne soit pas correct techniquement.

Maintenant, nous esquissons une démonstration de normalité asymptotique de la MLE de Type 2. Nous commençons par le développement de Taylor des équations de vraisemblance (8.12) autour de  $\theta_0$ , pour obtenir

$$0 = g(\hat{\theta}) = g(\theta_0) + H(\bar{\theta})(\hat{\theta} - \theta_0), \tag{8.35}$$

où  $\bar{\theta}$  est une combinaison convexe de  $\theta_0$  et  $\hat{\theta}$ , qui peut être différente pour chaque ligne de l'équation vectorielle. Si nous résolvons (8.35) par rapport à

$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$  et si nous récrivons tous les facteurs de manière à les rendre  $O(1)$ , nous obtenons

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -(n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}}))^{-1}(n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0)), \quad (8.36)$$

dans laquelle nous voyons que  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  égale une matrice de dimension  $k \times k$  fois un vecteur de dimension  $k$ . La matrice s'avèrera être asymptotiquement non aléatoire, et le vecteur s'avèrera être asymptotiquement normal, ce qui implique que  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  doit être asymptotiquement normal.

Nous voulons en premier lieu montrer que  $n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}})$  tend vers une certaine matrice limite non stochastique quand  $n \rightarrow \infty$ . Souvenons-nous que le  $ij^{\text{ième}}$  élément de  $n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}})$  est

$$\frac{1}{n} \sum_{t=1}^n \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \quad (8.37)$$

évalué en  $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ . Nous ferons en sorte que la condition WULLN s'applique à la série dont l'élément type est (8.37). Pour que cela soit réalisable,  $n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}})$  doit tendre vers  $\mathcal{H}(\bar{\boldsymbol{\theta}})$  quand  $n \rightarrow \infty$ . Mais comme  $\hat{\boldsymbol{\theta}}$  est convergent pour  $\boldsymbol{\theta}_0$  et que  $\bar{\boldsymbol{\theta}}$  reste entre  $\hat{\boldsymbol{\theta}}$  et  $\boldsymbol{\theta}_0$ , il est clair que  $n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}})$  doit également tendre vers  $\mathcal{H}(\boldsymbol{\theta}_0)$ . De plus, si le modèle est fortement asymptotiquement identifié, la matrice  $\mathcal{H}(\boldsymbol{\theta}_0)$  doit être définie négative, et nous supposerons que c'est effectivement le cas.

En utilisant cet argument et (8.36), nous voyons que

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\mathcal{H}^{-1}(\boldsymbol{\theta}_0)(n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0)). \quad (8.38)$$

Le seul élément stochastique dans le membre de droite de (8.38) est

$$n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0), \quad (8.39)$$

dont un élément type est

$$n^{-1/2} \sum_{t=1}^n \frac{\partial \log L_t(y_t, \boldsymbol{\theta})}{\partial \theta_i} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = n^{-1/2} \sum_{t=1}^n G_{ti}(\boldsymbol{\theta}_0).$$

Ainsi (8.39) est  $n^{-1/2}$  fois une somme de  $n$  quantités. D'après le résultat (8.33), nous savons que chacune de ces quantités a une espérance égale à zéro. Il semble alors plausible qu'un théorème de la limite centrale s'y applique. Dans une démonstration formelle, on devrait commencer par les conditions de régularité appropriées et les utiliser pour démontrer qu'un CLT particulier s'applique en effet à (8.39), mais nous omettrons cette étape. Si nous supposons que (8.39) est asymptotiquement normal, il suit immédiatement de (8.38) que  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  doit l'être également.

La **matrice de covariance asymptotique** de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  est simplement l'espérance asymptotique de  $n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top$ . En utilisant (8.38), cette quantité est égale à

$$(-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)) \left( \frac{1}{n} E_0(\mathbf{g}(\boldsymbol{\theta}_0)\mathbf{g}^\top(\boldsymbol{\theta}_0)) \right) (-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)).$$

Un élément type de l'espérance dans le facteur central est

$$\frac{1}{n} E_0 \left( \left( \sum_{t=1}^n G_{ti}(\boldsymbol{\theta}_0) \right) \left( \sum_{s=1}^n G_{sj}(\boldsymbol{\theta}_0) \right) \right). \quad (8.40)$$

Ceci est  $n^{-1}$  fois l'espérance du produit de deux sommes. Si nous devons développer explicitement le produit, nous verrions que chacun des termes dans la sommation des  $n^2$  termes dans (8.40) serait de la forme

$$G_{ti}(\boldsymbol{\theta}_0) G_{sj}(\boldsymbol{\theta}_0) = \frac{\partial \log(L_t)}{\partial \theta_i} \frac{\partial \log(L_s)}{\partial \theta_j}.$$

Tous ces termes doivent avoir une espérance égale à zéro, sauf quand  $t = s$ . Supposons sans perte de généralité que  $t > s$ . Alors

$$\begin{aligned} E_0(G_{ti}(\boldsymbol{\theta}_0) G_{sj}(\boldsymbol{\theta}_0)) &= E_0(E(G_{ti}(\boldsymbol{\theta}_0) G_{sj}(\boldsymbol{\theta}_0) | \mathbf{y}^s)) \\ &= E_0(G_{sj}(\boldsymbol{\theta}_0) E(G_{ti}(\boldsymbol{\theta}_0) | \mathbf{y}^s)) = 0. \end{aligned}$$

La dernière égalité provient du fait que  $E_0(G_{ti}(\boldsymbol{\theta}_0) | \mathbf{y}^s) = 0$ , qui est elle-même vraie parce que la preuve du résultat général (8.33) s'applique aussi bien à l'espérance conditionnelle qu'à l'espérance non conditionnelle.

Comme  $E_0(G_{ti}(\boldsymbol{\theta}_0) G_{sj}(\boldsymbol{\theta}_0)) = 0$  pour tout  $t \neq s$ ,

$$\frac{1}{n} E_0 \left( \left( \sum_{t=1}^n G_{ti}(\boldsymbol{\theta}_0) \right) \left( \sum_{s=1}^n G_{sj}(\boldsymbol{\theta}_0) \right) \right) = \frac{1}{n} E_0 \left( \sum_{t=1}^n G_{ti}(\boldsymbol{\theta}_0) G_{tj}(\boldsymbol{\theta}_0) \right). \quad (8.41)$$

De (8.20) et (8.21) nous voyons que le membre de droite de (8.41) correspond simplement à  $\mathcal{J}^n(\boldsymbol{\theta}_0)$ , la matrice d'information moyenne pour un échantillon de taille  $n$ . En utilisant le fait que  $\mathcal{J}(\boldsymbol{\theta}_0)$  est la limite de  $\mathcal{J}^n(\boldsymbol{\theta}_0)$  quand  $n \rightarrow \infty$ , nous concluons que la matrice de covariance asymptotique de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  est

$$\mathbf{V}^\infty(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) = \mathcal{H}^{-1}(\boldsymbol{\theta}_0) \mathcal{J}(\boldsymbol{\theta}_0) \mathcal{H}^{-1}(\boldsymbol{\theta}_0). \quad (8.42)$$

Dans la prochaine section, nous verrons que cette expression peut être grandement simplifiée.

Nous pouvons à présent établir les résultats précédents comme suit:

*Théorème 8.3. Théorème de Normalité Asymptotique.*

L'estimateur ML de Type 2,  $\hat{\boldsymbol{\theta}}$ , pour un modèle fortement identifié asymptotiquement représenté par la famille paramétrique des fonctions de logvraisemblance  $\ell(\boldsymbol{\theta})$ ,  $\boldsymbol{\theta} \in \Theta$ , quand il existe et est convergent pour le vecteur paramétrique  $\boldsymbol{\theta}_0$  qui caractérise le DGP, est asymptotiquement normal si

- (i) les contributions  $\ell_t(\mathbf{y}, \boldsymbol{\theta})$  à  $\ell$  sont au moins deux fois continûment différentiables en  $\boldsymbol{\theta}$  pour presque tout  $\mathbf{y}$  et tout  $\boldsymbol{\theta} \in \Theta$ ,
- (ii) les séries composantes de  $\{D_{\theta\theta}^2 \ell_t(\mathbf{y}, \boldsymbol{\theta})\}_{t=1}^{\infty}$  satisfont la condition WULLN sur  $\Theta$ , et
- (iii) les séries composantes de  $\{D_{\theta} \ell_t(\mathbf{y}, \boldsymbol{\theta})\}_{t=1}^{\infty}$  satisfont la condition CLT.

Par le terme de normalité asymptotique, nous signifions que la série de variables aléatoires  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  a une limite en probabilité qui est une variable aléatoire de l'ordre de l'unité, normalement distribuée d'espérance nulle et de matrice de covariance (8.42).

## 8.6 L'ÉGALITÉ DE LA MATRICE D'INFORMATION

Dans cette section, nous établirons un résultat important qui permet une simplification substantielle de l'expression (8.42) de la matrice de covariance asymptotique de l'estimateur ML. Ce résultat, qui, comme l'annonce le titre de la section, est connu sous le nom de **l'égalité de la matrice d'information**, est

$$\mathcal{H}(\boldsymbol{\theta}_0) = -\mathcal{I}(\boldsymbol{\theta}_0). \quad (8.43)$$

Littéralement, la matrice d'information Hessienne asymptotique est l'opposé de la matrice d'information asymptotique. Un résultat analogue est vrai pour des observations individuelles:

$$E_0(D_{\theta\theta}^2 \ell_t(\mathbf{y}, \boldsymbol{\theta}_0)) = -E_0(D_{\theta}^{\top} \ell_t(\mathbf{y}, \boldsymbol{\theta}_0) D_{\theta} \ell_t(\mathbf{y}, \boldsymbol{\theta}_0)). \quad (8.44)$$

Le dernier résultat implique clairement le premier, étant données les hypothèses qui permettent l'application d'une loi des grands nombres aux séries  $\{D_{\theta\theta}^2 \ell_t(\mathbf{y}, \boldsymbol{\theta}_0)\}_{t=1}^{\infty}$  et  $\{D_{\theta}^{\top} \ell_t(\mathbf{y}, \boldsymbol{\theta}_0) D_{\theta} \ell_t(\mathbf{y}, \boldsymbol{\theta}_0)\}_{t=1}^{\infty}$ .

Le résultat (8.44) est démontré à l'aide d'un argument très similaire à celui utilisé au début de la dernière section pour montrer que l'espérance de la matrice CG est égale zéro. Du fait que

$$\frac{\partial \ell_t}{\partial \theta_i} = \frac{1}{L_t} \frac{\partial L_t}{\partial \theta_i},$$

nous obtenons après une différentiation supplémentaire:

$$\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} = \frac{1}{L_t} \frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j} - \frac{1}{L_t^2} \frac{\partial L_t}{\partial \theta_i} \frac{\partial L_t}{\partial \theta_j}.$$

En conséquence,

$$\frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} = \frac{1}{L_t} \frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j}. \quad (8.45)$$

Maintenant, si nous calculons l'espérance de (8.45) pour le DGP caractérisé par la même valeur du vecteur paramétrique  $\boldsymbol{\theta}$  que celle avec laquelle les fonctions  $\ell_t$  et  $L_t$  sont évaluées (que nous désignerons comme d'habitude par  $E_\theta$ ), nous trouvons que

$$\begin{aligned} E_\theta \left( \frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} \right) &= \int L_t \frac{1}{L_t} \frac{\partial^2 L_t}{\partial \theta_i \partial \theta_j} dy_t \\ &= \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int L_t dy_t = 0, \end{aligned} \quad (8.46)$$

à condition que, comme pour (8.34), la permutation de l'ordre de différentiation et d'intégration puisse être justifiée. Alors, le résultat (8.46) établit (8.44), puisqu'il implique que

$$E_\theta \left( \frac{\partial^2 \ell_t}{\partial \theta_i \partial \theta_j} \right) = 0 - E_\theta \left( \frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} \right) = -E_\theta \left( \frac{\partial \ell_t}{\partial \theta_i} \frac{\partial \ell_t}{\partial \theta_j} \right).$$

Afin d'établir (8.43), rappelons que, à partir de (8.19) et de la loi des grands nombres,

$$\begin{aligned} \mathcal{H}(\boldsymbol{\theta}) &= \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n E_\theta \left( \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \right) \\ &= - \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n E_\theta \left( \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_j} \right) \right) \\ &= -\mathcal{J}(\boldsymbol{\theta}), \end{aligned}$$

où la dernière ligne provient directement de la définition de la matrice d'information asymptotique, (8.22). Alors ceci donne (8.43).

En substituant soit  $-\mathcal{H}(\boldsymbol{\theta}_0)$  à  $\mathcal{J}(\boldsymbol{\theta}_0)$  soit  $\mathcal{J}(\boldsymbol{\theta}_0)$  à  $-\mathcal{H}(\boldsymbol{\theta}_0)$  dans (8.42), il est facile de conclure que la matrice de covariance asymptotique de l'estimateur ML est donnée par l'une ou l'autre des deux expressions équivalentes  $-\mathcal{H}(\boldsymbol{\theta}_0)^{-1}$  et  $\mathcal{J}(\boldsymbol{\theta}_0)^{-1}$ . Formellement, nous pouvons écrire

$$\mathbf{V}^\infty(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) = \mathcal{J}^{-1}(\boldsymbol{\theta}_0) = -\mathcal{H}^{-1}(\boldsymbol{\theta}_0).$$

Afin d'effectuer une quelconque inférence statistique, il est nécessaire de pouvoir *estimer*  $\mathcal{J}^{-1}(\boldsymbol{\theta}_0)$  ou  $-\mathcal{H}^{-1}(\boldsymbol{\theta}_0)$ . L'estimateur qui vient immédiatement à l'esprit est  $\mathcal{J}^{-1}(\hat{\boldsymbol{\theta}})$ , c'est-à-dire l'inverse de la matrice d'information asymptotique évaluée avec la MLE,  $\hat{\boldsymbol{\theta}}$ . Notons que la fonction matricielle  $\mathcal{J}(\boldsymbol{\theta})$  *n'est pas* un objet dépendant de l'échantillon. Elle peut, en principe, être calculée théoriquement comme une fonction matricielle des paramètres du modèle à partir (de la série) des fonctions de logvraisemblance  $\ell^n$ . Pour certains modèles, c'est un calcul entièrement réalisable, et cela donne alors ce qui est souvent l'estimateur préféré de la matrice de covariance asymptotique.

Mais pour certains modèles, le calcul, même s'il était réalisable, serait excessivement laborieux, et dans ces cas, il est commode de disposer d'autres estimateurs convergents de  $\mathcal{J}(\boldsymbol{\theta}_0)$  et en conséquence de la matrice de covariance asymptotique.

Un estimateur commun est l'opposé de ce que l'on nomme **matrice Hessienne empirique**. Cette matrice est définie comme

$$\hat{\mathcal{H}} \equiv \mathcal{H}^n(\mathbf{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{t=1}^n D_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 \ell_t(\mathbf{y}, \hat{\boldsymbol{\theta}}); \quad (8.47)$$

elle correspond simplement à  $\mathcal{H}^n(\mathbf{y}, \boldsymbol{\theta})$  évaluée en  $\hat{\boldsymbol{\theta}}$ . La loi des grands nombres et la convergence de  $\hat{\boldsymbol{\theta}}$  elle-même garantissent immédiatement la convergence de (8.47) pour  $\mathcal{H}(\boldsymbol{\theta}_0)$ . Quand la matrice Hessienne empirique est directement disponible, comme cela sera le cas si les programmes de maximisation qui utilisent les dérivées secondes sont employés, l'opposé de son inverse peut fournir une manière très commode d'estimer la matrice de covariance de  $\boldsymbol{\theta}$ . Cependant, la matrice Hessienne est souvent difficile à calculer, et si elle n'est pas déjà calculée pour d'autres fins, il est probablement insensé de la calculer uniquement pour estimer une matrice de covariance.

Un autre estimateur de la matrice de covariance communément utilisé est connu sous le nom d'**estimateur produit-extérieur-du-gradient**, ou **estimateur OPG**. Il est basé sur la définition

$$\mathcal{J}(\boldsymbol{\theta}) \equiv \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n E_{\boldsymbol{\theta}} (D_{\boldsymbol{\theta}}^{\top} \ell_t(\boldsymbol{\theta}) D_{\boldsymbol{\theta}} \ell_t(\boldsymbol{\theta})) \right).$$

L'estimateur OPG est

$$\hat{\mathcal{J}}_{\text{OPG}} \equiv \frac{1}{n} \sum_{t=1}^n D_{\boldsymbol{\theta}}^{\top} \ell_t(\mathbf{y}, \hat{\boldsymbol{\theta}}) D_{\boldsymbol{\theta}} \ell_t(\mathbf{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \mathbf{G}^{\top}(\hat{\boldsymbol{\theta}}) \mathbf{G}(\hat{\boldsymbol{\theta}}), \quad (8.48)$$

et sa convergence est garantie une fois de plus par la condition CLT, qui inclut une loi des grands nombres pour la somme dans (8.48).

L'estimateur OPG de la matrice d'information a été préconisé par Berndt, Hall, Hall, et Hausman (1974) dans un article célèbre et on s'y réfère parfois sous le nom de l'estimateur BHHH. Ils ont aussi suggéré son utilisation comme partie d'un système général pour la maximisation de fonctions de logvraisemblance, analogue aux systèmes basés sur la régression de Gauss-Newton dont nous avons discuté dans la Section 6.8. Malheureusement, l'estimateur (8.48) passe pour être plutôt bruité dans la pratique, ce qui limite son utilité à l'estimation des matrices de covariance et à la maximisation numérique.<sup>7</sup>

<sup>7</sup> Il y aura quelques discussions supplémentaires dans le Chapitre 13 sur les manières alternatives d'estimer la matrice de covariance. Pour une discussion de la performance de l'estimateur OPG dans le système d'estimation BHHH, consulter Belsley (1980).



Alors que dans  $\mathcal{J}(\hat{\boldsymbol{\theta}})$  le seul élément stochastique est la MLE  $\hat{\boldsymbol{\theta}}$  elle-même, à la fois la matrice Hessienne empirique et l'estimateur OPG dépendent explicitement de l'échantillon réalisé  $\mathbf{y}$ , et cette dépendance leur transmet un bruit additionnel qui rend les inférences basées sur ces estimateurs moins fiables que l'on ne le souhaiterait. Souvent l'estimateur OPG semble être particulièrement pauvre, comme nous en discuterons dans le Chapitre 13.

Dans certains cas, il est possible de trouver des estimateurs quelque part entre l'estimateur (habituellement) préféré  $\mathcal{J}(\hat{\boldsymbol{\theta}})$  et l'estimateur OPG, pour lequel on peut calculer les espérances de certains des termes apparaissant dans (8.48) mais pas de tous. Ceci semble être une bonne procédure à suivre pour la qualité de l'inférence statistique que l'on peut obtenir à partir des distributions asymptotiques des estimateurs ou des statistiques de test. L'estimateur Gauss-Newton de la matrice de covariance est de ce type chaque fois que le modèle contient des variables dépendantes retardées, car la matrice  $n^{-1}\mathbf{X}^\top(\hat{\boldsymbol{\beta}})\mathbf{X}(\hat{\boldsymbol{\beta}})$  dépendra alors de valeurs retardées de  $\mathbf{y}$  aussi bien que de  $\hat{\boldsymbol{\beta}}$ . Beaucoup plus d'exemples de ce type d'estimateur apparaîtront plus tard dans ce livre, plus particulièrement dans les Chapitres 14 et 15.

La discussion précédente n'a peut-être pas rendu clair un point qui est de la plus haute importance pratique quand on essaie de pratiquer des inférences concernant un ensemble d'estimations ML  $\hat{\boldsymbol{\theta}}$ . Tout ce qui se rattache à la théorie de la distribution asymptotique se note en terme de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ , mais en pratique nous voulons en fait utiliser  $\hat{\boldsymbol{\theta}}$  pour réaliser des inférences à propos de  $\boldsymbol{\theta}$ . Ceci signifie que nous devons baser nos inférences *non pas* sur des quantités qui estiment  $\mathcal{J}(\boldsymbol{\theta}_0)$  mais plutôt sur des quantités qui estiment  $n\mathcal{J}(\boldsymbol{\theta}_0)$ . Alors les trois estimateurs qui peuvent être utilisés en pratique pour estimer  $\mathbf{V}(\hat{\boldsymbol{\theta}})$  sont l'inverse de l'opposé de la matrice Hessienne numérique,

$$(-\mathbf{H}(\hat{\boldsymbol{\theta}}))^{-1}, \quad (8.49)$$

l'inverse de l'estimateur OPG de la matrice d'information  $O(n)$ ,

$$(\mathbf{G}^\top(\hat{\boldsymbol{\theta}})\mathbf{G}(\hat{\boldsymbol{\theta}}))^{-1}, \quad (8.50)$$

et l'inverse de la matrice d'information  $O(n)$  elle-même,

$$(n\mathcal{J}(\hat{\boldsymbol{\theta}}))^{-1} \equiv (\mathbf{I}^n(\hat{\boldsymbol{\theta}}))^{-1}. \quad (8.51)$$

En plus de (8.49), (8.50) et (8.51), qui sont très largement applicables, il y a des estimateurs hybrides variés pour certaines classes de modèles, tels que les estimateurs basés sur les régressions de Gauss-Newton et sur d'autres régressions artificielles. Notons que tous ces estimateurs de matrice de covariance seront  $n$  fois plus petits que les estimateurs de la matrice de covariance  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ , tels que (8.47) et (8.48), dont nous avons discuté jusqu'ici.

Bien qu'il soit commun de calculer autant d'espérances que possible quand on estime la matrice de covariance de  $\hat{\boldsymbol{\theta}}$ , il n'est pas évident que cela

soit toujours une bonne chose. Considérons l'exemple suivant. Supposons que  $y_t = \beta x_t + u_t$ , où  $x_t$  est une variable binaire dont nous savons qu'elle prend la valeur 1 avec une probabilité  $p$  et la valeur 0 avec la probabilité  $1 - p$ . Supposons de plus (pour simplifier) que la variance de  $u_t$  soit connue et égale à l'unité. Alors la matrice d'information, qui est simplement un scalaire dans ce cas, est  $E(n^{-1} \sum_{t=1}^n x_t^2) = p$ . Ainsi l'estimation usuelle de la variance de  $\hat{\beta}$  basée sur la matrice d'information est simplement  $(np)^{-1}$ .

Il devrait être évident que, quand  $np$  est petit,  $(np)^{-1}$  pourrait être une estimation très trompeuse de la variance réelle de  $\hat{\beta}$  conditionnelle à l'échantillon particulier qui a été observé. Supposons, par exemple, que  $n$  soit 100 et  $p$  soit .02. L'estimation habituelle de la variance serait  $\frac{1}{2}$ . Mais il pourrait survenir qu'aucun des  $x_t$  de l'échantillon ne soit égal à 1; ceci arriverait avec une probabilité .133. Alors cet échantillon particulier n'identifierait pas du tout  $\beta$ , et la variance de  $\hat{\beta}$  serait infinie. De façon contraire, il peut survenir qu'un seul des  $x_t$  dans l'échantillon soit égal à 1. Alors  $\beta$  serait identifié, mais  $\frac{1}{2}$  serait à l'évidence une sous-estimation de la variance réelle de  $\hat{\beta}$ . D'un autre côté, si plus de deux des  $x_t$  étaient égaux à 1,  $\hat{\beta}$  aurait une variance plus petite que  $(np)^{-1}$ . L'estimation de la variance asymptotique ne correspondrait à la véritable variance de  $\hat{\beta}$  conditionnelle à l'échantillon observé que dans le cas où  $np$  était égal à sa valeur espérée, 2.

Cet exemple est très spécial, mais le phénomène qu'il illustre est assez général. A chaque fois que nous calculons la matrice de covariance d'un certain vecteur d'estimations paramétriques, nous nous soucions vraisemblablement de la précision de cet ensemble particulier d'estimations. Cela dépend de la quantité d'information qui a été fournie par l'échantillon dont nous disposons plutôt que de la quantité d'information qui serait fournie par un échantillon type de la même taille. Désormais, dans un sens très concret, c'est la matrice d'information *observée* plutôt que la matrice d'information *attendue* qui devrait nous intéresser. Pour une discussion beaucoup plus étendue sur ce point, consulter Efron et Hinkley (1978).

## 8.7 LA FONCTION DE LOGVRAISEMBLANCE CONCENTRÉE

Il arrive souvent que les paramètres dont dépend une fonction de logvraisemblance puissent être partitionnés en deux ensembles de façon à rendre facile l'écriture de l'estimateur ML d'un groupe de paramètres comme une fonction des valeurs de l'autre groupe. Nous rencontrerons un exemple de ceci, en connexion avec l'estimation ML des modèles de régression, dans la Section 8.10, et d'autres exemples dans le Chapitre 9. Dans cette situation, il peut être très pratique de **concentrer** la fonction de logvraisemblance en l'écrivant comme une fonction d'un seul des deux groupes de paramètres. Supposons que nous puissions écrire la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$  comme  $\ell(\mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ . Les conditions du premier ordre qui définissent les estimateurs ML (de Type 2)

$\hat{\theta}_1$  et  $\hat{\theta}_2$  sont

$$D_1\ell(\mathbf{y}, \theta_1, \theta_2) = \mathbf{0} \quad \text{et} \quad D_2\ell(\mathbf{y}, \theta_1, \theta_2) = \mathbf{0},$$

où, comme d'habitude,  $D_i\ell$  désigne le vecteur ligne des dérivées partielles  $\partial\ell/\partial\theta_i$  pour  $i = 1, 2$ . Supposons qu'il soit possible de résoudre le second ensemble de conditions du premier ordre, afin de pouvoir écrire

$$\theta_2 = \tau(\mathbf{y}, \theta_1).$$

Ceci implique alors que, identiquement en  $\theta_1$ ,

$$D_2\ell(\mathbf{y}, \theta_1, \tau(\mathbf{y}, \theta_1)) = \mathbf{0}. \quad (8.52)$$

En substituant  $\tau(\mathbf{y}, \theta_1)$  à  $\theta_2$  dans  $\ell(\mathbf{y}, \theta_1, \theta_2)$ , nous obtenons la **fonction de logvraisemblance concentrée**

$$\ell^c(\mathbf{y}, \theta_1) \equiv \ell(\mathbf{y}, \theta_1, \tau(\mathbf{y}, \theta_1)).$$

Si  $\hat{\theta}_1$  maximise celle-ci, nous pouvons alors obtenir  $\hat{\theta}_2$  grâce à  $\tau(\mathbf{y}, \hat{\theta}_1)$ , et il est évident que  $[\hat{\theta}_1; \hat{\theta}_2]$  maximisera  $\ell(\mathbf{y}, \theta)$ . Dans certains cas, cette stratégie peut réduire substantiellement la quantité d'efforts nécessaires à l'obtention des estimations ML.

Il est évident que  $\ell^c(\mathbf{y}, \hat{\theta}_1)$  sera identique à  $\ell(\mathbf{y}, \hat{\theta})$ . Cependant, il n'est pas évident que nous puissions calculer une matrice de covariance estimée pour  $\hat{\theta}_1$  basée sur  $\ell^c(\mathbf{y}, \theta_1)$  de la même manière que celle que nous calculons lorsque nous nous basons sur  $\ell(\mathbf{y}, \theta)$ . En fait, à condition d'utiliser comme estimateur l'inverse de l'opposée de la matrice Hessienne empirique, on dispose d'un estimateur évident. La raison est que, en vertu de la manière dont  $\ell^c$  est construite, l'inverse de sa matrice Hessienne par rapport à  $\theta_1$  est égale au bloc  $(\theta_1, \theta_1)$  de l'inverse de la matrice Hessienne de  $\ell(\mathbf{y}, \theta)$  par rapport au vecteur paramétrique entier  $\theta$ . Ceci provient du théorème de l'enveloppe et des résultats standards sur les matrices partitionnées, comme nous allons le démontrer à présent.

Grâce aux conditions du premier ordre (8.52), le gradient de  $\ell^c$  par rapport à  $\theta_1$  est

$$\begin{aligned} D_1\ell^c(\theta_1) &= D_1\ell(\theta_1, \tau(\theta_1)) + D_2\ell(\theta_1, \tau(\theta_1))D\tau(\theta_1) \\ &= D_1\ell(\theta_1, \tau(\theta_1)), \end{aligned}$$

où la dépendance explicite à  $\mathbf{y}$  a été supprimée. Ce résultat est simplement le théorème de l'enveloppe appliqué à  $\ell^c$ . Ainsi la matrice Hessienne de  $\ell^c(\theta_1)$  est

$$D_{11}\ell^c(\theta_1) = D_{11}\ell(\theta_1, \tau(\theta_1)) + D_{12}\ell(\theta_1, \tau(\theta_1))D\tau(\theta_1). \quad (8.53)$$

Afin d'exprimer le membre de droite de (8.53) en termes uniquement des blocs de la matrice Hessienne de  $\ell$ , nous dérivons (8.52) par rapport à  $\boldsymbol{\theta}_1$ , et obtenons

$$D_{21}\ell(\boldsymbol{\theta}_1, \boldsymbol{\tau}(\boldsymbol{\theta}_1)) + D_{22}\ell(\boldsymbol{\theta}_1, \boldsymbol{\tau}(\boldsymbol{\theta}_1))D\boldsymbol{\tau}(\boldsymbol{\theta}_1) = \mathbf{0}.$$

En résolvant cette équation pour  $D\boldsymbol{\tau}(\boldsymbol{\theta}_1)$  et en substituant le résultat dans (8.53), l'expression de la matrice Hessienne de  $\ell^c$ , nous aboutissons à

$$D_{11}\ell^c = D_{11}\ell - D_{12}\ell(D_{22}\ell)^{-1}D_{21}\ell, \quad (8.54)$$

expression dans laquelle les arguments de  $\ell$  et  $\ell^c$  ont été omis pour simplifier l'écriture. La matrice Hessienne de  $\ell$  peut être écrite sous forme partitionnée comme

$$D_{\theta\theta}\ell = \begin{bmatrix} D_{11}\ell & D_{12}\ell \\ D_{21}\ell & D_{22}\ell \end{bmatrix}.$$

Les résultats standards sur les matrices partitionnées (consulter l'Annexe A) nous apprennent que le bloc  $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)$  de l'inverse de cette matrice Hessienne est

$$(D_{11}\ell - D_{12}\ell(D_{22}\ell)^{-1}D_{21}\ell)^{-1},$$

dont l'inverse est précisément l'expression pour  $D_{11}\ell^c$  dans (8.54).

L'utilisation des fonctions de logvraisemblance concentrées comporte certains désavantages. La fonction de logvraisemblance originelle peut dans la plupart des cas être écrite de manière commode comme

$$\ell(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^n \ell_t(y_t, \boldsymbol{\theta}). \quad (8.55)$$

Ceci n'est cependant généralement pas exact pour la fonction de logvraisemblance concentrée. L'équivalent de (8.55) est

$$\ell^c(\mathbf{y}, \boldsymbol{\theta}_1) = \sum_{t=1}^n \ell_t(y_t, \boldsymbol{\theta}_1, \boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}_1)),$$

et il est évident qu'en raison de la dépendance de  $\boldsymbol{\tau}(\cdot)$  au vecteur entier  $\mathbf{y}$ , il n'y a pas en général de manière simple d'écrire  $\ell^c(\mathbf{y}, \boldsymbol{\theta}_1)$  comme une somme des contributions de chacune des observations. Cela signifie que l'estimateur OPG de la matrice d'information n'est généralement pas disponible pour les fonctions de logvraisemblance concentrées. On peut bien sûr utiliser  $\ell^c(\mathbf{y}, \boldsymbol{\theta}_1)$  pour l'estimation et se reporter ensuite vers  $\ell(\mathbf{y}, \boldsymbol{\theta})$  quand vient l'heure d'estimer la matrice de covariance des estimations.

## 8.8 L'EFFICACITÉ ASYMPTOTIQUE DE L'ESTIMATEUR ML

Dans cette section, nous démontrerons l'**efficacité asymptotique** de l'estimateur ML ou, à proprement parler, de l'estimateur ML de Type 2. La convergence asymptotique signifie que la variance de la distribution asymptotique de n'importe quel estimateur convergent des paramètres diffère de celle d'un estimateur efficace asymptotiquement par une matrice semi-définie positive; voir la Définition 5.6. On parle d'*un* estimateur efficace asymptotiquement plutôt que de *l'*estimateur efficace asymptotiquement parce que la propriété d'efficacité asymptotique est une propriété de la distribution asymptotique seulement; il peut exister de nombreux estimateurs (et il en existera effectivement) qui diffèrent avec des échantillons finis mais qui ont la même distribution asymptotique efficace. Un exemple de modèle de régression non linéaire peut être pris, dans lequel, comme nous le verrons dans la Section 8.10, l'estimation NLS est équivalente à l'estimation ML si nous supposons la normalité des aléas. Comme nous l'avons vu dans la Section 6.6, il existe des modèles non linéaires qui correspondent exactement à des modèles linéaires auxquels on impose certaines contraintes non linéaires. Dans de tels cas nous avons vu que l'estimation en une étape qui commence à partir des estimations de modèle linéaire était asymptotiquement équivalente à l'estimation NLS, et par conséquent asymptotiquement efficace. L'estimation en une étape est aussi possible dans le contexte général du maximum de vraisemblance et peut souvent fournir un estimateur efficace qui est plus facile à calculer que l'estimateur ML lui-même.

Nous commençons notre démonstration de l'efficacité asymptotique de l'estimateur ML par une discussion applicable à n'importe quel estimateur convergent, au taux  $n^{1/2}$  et asymptotiquement sans biais, des paramètres du modèle représenté par la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$ . Notons que la convergence en elle-même n'implique pas l'absence de biais asymptotiquement sans l'imposition de diverses conditions de régularité. Puisque tout estimateur convergent et intéressant au sens économétrique que nous connaissons est en fait asymptotiquement sans biais, nous ne traiterons ici que de tels estimateurs. Désignons un tel estimateur par  $\hat{\boldsymbol{\theta}}(\mathbf{y})$ , avec une notation qui insiste sur le fait que l'estimateur est une variable aléatoire qui dépend de l'échantillon  $\mathbf{y}$  réalisé. Notons que nous avons changé ici de notation, car  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  n'est pas en général l'estimateur ML. Au lieu de cela, ce dernier sera noté  $\boldsymbol{\theta}(\mathbf{y})$ ; la nouvelle notation est conçue pour être cohérente, à travers l'ouvrage, avec notre traitement des estimateurs contraints et non contraints, puisque dans un sens profond l'estimateur ML correspond aux premiers de ces estimateurs et l'estimateur convergent arbitraire  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  correspond aux seconds.

Comme  $\hat{\boldsymbol{\theta}}(\mathbf{y})$  est supposé être asymptotiquement sans biais, nous avons

$$\lim_{n \rightarrow \infty} E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}) = \mathbf{0}.$$

Avec une notation plus explicite, ceci devient:

$$\lim_{n \rightarrow \infty} \left( \int_{\mathcal{Y}^n} L^n(\mathbf{y}^n, \boldsymbol{\theta}) \hat{\boldsymbol{\theta}}^n(\mathbf{y}^n) d\mathbf{y}^n - \boldsymbol{\theta} \right) = \mathbf{0}, \quad (8.56)$$

où, comme précédemment,  $\mathcal{Y}^n$  désigne le sous-espace de  $\mathbb{R}^{nm}$  sur lequel le vecteur échantillon  $\mathbf{y}^n$  peut varier en conservant une taille  $n$ . Les prochaines étapes impliquent la différentiation de la relation (8.56) par rapport aux éléments de  $\boldsymbol{\theta}$ , en permutant l'ordre des opérations de différentiation et d'intégration, et en calculant la limite quand  $n \rightarrow \infty$ . Nous omettons la discussion sur les conditions de régularité nécessaires pour que ceci soit admissible et poursuivons en écrivant directement le résultat de la différentiation du  $j^{\text{ième}}$  élément de (8.56) par rapport au  $i^{\text{ième}}$  élément de  $\boldsymbol{\theta}$ :

$$\lim_{n \rightarrow \infty} \int_{\mathcal{Y}^n} L^n(\mathbf{y}^n, \boldsymbol{\theta}) \frac{\partial \ell^n(\mathbf{y}^n, \boldsymbol{\theta})}{\partial \theta_i} \hat{\theta}_j(\mathbf{y}^n) d\mathbf{y}^n = \delta_j^i. \quad (8.57)$$

Le membre de droite de cette équation est le delta de Kronecker, égal à 1 quand  $i = j$  et égal à 0 sinon. L'équation (8.57) peut être réécrite comme

$$\lim_{n \rightarrow \infty} E_{\boldsymbol{\theta}} \left( n^{-1/2} \frac{\partial \ell^n(\mathbf{y}^n, \boldsymbol{\theta})}{\partial \theta_i} n^{1/2} (\hat{\theta}_j - \theta_j) \right) = \delta_j^i, \quad (8.58)$$

où nous avons introduit certaines puissances de  $n$  pour s'assurer que les quantités qui apparaissent dans l'expression possèdent des limites en probabilité de l'ordre de l'unité. Nous avons aussi retranché  $\theta_j$  à  $\hat{\theta}_j$ ; ceci a été possible parce que  $E_{\boldsymbol{\theta}}(D_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})) = \mathbf{0}$ , et désormais le produit de  $\hat{\theta}_j$  par  $E_{\boldsymbol{\theta}}(D_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}))$  est également nul.

L'expression (8.58) peut être écrite sans aucune opération à la limite si nous utilisons les distributions asymptotiques du gradient  $D_{\boldsymbol{\theta}} \ell$  et le vecteur  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ . Introduisons une notation supplémentaire dans le but de discuter des variables aléatoires asymptotiques. Nous posons les définitions

$$\mathbf{s}^n(\boldsymbol{\theta}) \equiv n^{-1/2} \mathbf{g}(\mathbf{y}^n, \boldsymbol{\theta}), \quad \mathbf{s}(\boldsymbol{\theta}) \equiv \text{plim}_{n \rightarrow \infty} \mathbf{s}^n(\boldsymbol{\theta}), \quad (8.59)$$

$$\hat{\mathbf{t}}^n(\boldsymbol{\theta}) \equiv n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \quad \text{et} \quad \hat{\mathbf{t}}(\boldsymbol{\theta}) \equiv \text{plim}_{n \rightarrow \infty} \hat{\mathbf{t}}^n(\boldsymbol{\theta}). \quad (8.60)$$

Ainsi  $\mathbf{s}(\boldsymbol{\theta})$  et  $\hat{\mathbf{t}}(\boldsymbol{\theta})$  sont des vecteurs de dimension  $k$  dont les éléments types respectifs sont  $s_i(\boldsymbol{\theta})$  et  $\hat{t}_j(\boldsymbol{\theta})$ . Le premier est la valeur à la limite de  $n^{-1/2}$  fois un élément type du gradient de  $\ell(\mathbf{y}, \boldsymbol{\theta})$ , tandis que le second est la valeur à la limite de  $n^{1/2}$  fois un élément type de la différence entre  $\hat{\boldsymbol{\theta}}$  et  $\boldsymbol{\theta}$ . La notation a été conçue dans l'intention d'être mnémotechnique,  $\mathbf{s}(\boldsymbol{\theta})$  correspondant au vecteur *score* et  $\hat{\mathbf{t}}(\boldsymbol{\theta})$  correspondant au *thêta chapeau*. Grâce à cette nouvelle notation commode, l'expression (8.58) devient

$$E_{\boldsymbol{\theta}}(\hat{\mathbf{t}}(\boldsymbol{\theta}) \mathbf{s}^{\top}(\boldsymbol{\theta})) = \mathbf{I}_k, \quad (8.61)$$

où  $\mathbf{I}_k$  est simplement la matrice identité de dimension  $k \times k$ .

Il n'est pas en général exact pour *n'importe quel* estimateur convergent que la limite en probabilité dans (8.60) existe ou, si elle existe, qu'elle soit non nulle. La classe des estimateurs pour lesquels celle-ci existe et n'est pas nulle est appelée la classe des **estimateurs convergents au taux**  $n^{1/2}$ . Ainsi que nous en avons discuté dans le Chapitre 5, ceci signifie que le taux de convergence, quand  $n \rightarrow \infty$ , de l'estimateur  $\hat{\boldsymbol{\theta}}$  vers la véritable valeur  $\boldsymbol{\theta}$  est le même que le taux de convergence de  $n^{-1/2}$  vers zéro. L'existence d'une limite en probabilité non nulle dans (8.60) implique clairement cette propriété, et nous avons déjà montré que l'estimateur ML est convergent au taux  $n^{1/2}$ . La convergence de  $\hat{\boldsymbol{\theta}}$  implique également que l'espérance de la variable aléatoire à la limite  $\hat{\mathbf{t}}(\boldsymbol{\theta})$  est égale à zéro.

Pour la partie suivante de l'argumentation, nous considérons en premier lieu le cas simple dans lequel  $k = 1$ . Alors à la place de (8.61) nous avons la relation scalaire

$$E_{\theta}(\hat{t}(\theta)s(\theta)) = \text{Cov}_{\theta}(\hat{t}(\theta), s(\theta)) = 1. \quad (8.62)$$

Ici nous avons utilisé le fait que les espérances aussi bien de  $\hat{t}(\theta)$  que de  $s(\theta)$  sont zéro. Le résultat (8.62) implique l'inégalité bien connue de Cauchy-Schwartz:

$$1 = \left( \text{Cov}_{\theta}(\hat{t}(\theta), s(\theta)) \right)^2 \leq \text{Var}_{\theta}(\hat{t}(\theta)) \text{Var}_{\theta}(s(\theta)) = \text{Var}_{\theta}(\hat{t}(\theta)) \mathcal{J}(\theta), \quad (8.63)$$

où la dernière égalité provient de la définition (8.59) de  $s(\theta)$  et de la définition de la matrice d'information asymptotique  $\mathcal{J}(\theta)$ , qui est dans ce cas un scalaire. L'inégalité (8.63) implique que

$$\text{Var}_{\theta}(\hat{t}(\theta)) \geq \frac{1}{\mathcal{J}(\theta)}. \quad (8.64)$$

Ce résultat établit, dans ce cas à une dimension, que la variance asymptotique de n'importe quel estimateur convergent à un taux  $n^{1/2}$  ne peut pas être inférieure à l'inverse de ce qu'il semble être logique d'appeler le scalaire d'information. Comme le membre de droite de (8.64) est précisément la variance asymptotique de l'estimateur ML, l'efficacité asymptotique de ce dernier est aussi établie par ce résultat. Notons que (8.64) élimine n'importe quel estimateur pour lequel la limite en probabilité de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  est égale à zéro. Un tel estimateur serait naturellement *plus* efficace asymptotiquement que l'estimateur ML, car il devrait converger plus rapidement vers la véritable valeur de  $\boldsymbol{\theta}$ .

Le résultat général analogue à (8.64) pour le cas  $k \geq 1$  peut maintenant être établi en ajoutant un tout petit peu plus de travail. Considérons la matrice entière de covariance de tous les éléments de  $\hat{\mathbf{t}}$  et de  $\mathbf{s}$ , c'est-à-dire la

matrice de covariance de  $[\hat{\mathbf{t}}(\boldsymbol{\theta}) ; \mathbf{s}(\boldsymbol{\theta})]$ . Notons  $\mathbf{V}$  la matrice de covariance de  $\hat{\mathbf{t}}$ . Alors (8.61) et le fait que  $\text{Var}_{\boldsymbol{\theta}}(\mathbf{s}^{\top}(\boldsymbol{\theta})) = \mathcal{J}(\boldsymbol{\theta})$  signifient que la matrice de covariance de  $[\hat{\mathbf{t}}(\boldsymbol{\theta}) ; \mathbf{s}(\boldsymbol{\theta})]$  peut être écrite comme

$$\text{Var}(\hat{\mathbf{t}}, \mathbf{s}) = \begin{bmatrix} \mathbf{V} & \mathbf{I}_k \\ \mathbf{I}_k & \mathcal{J} \end{bmatrix}.$$

Comme il s'agit d'une matrice de covariance, celle-ci doit être semi-définie positive. Ainsi, pour n'importe quel vecteur  $\mathbf{a}$  de dimension  $k$ , l'expression suivante est non négative:

$$[\mathbf{a}^{\top} - \mathbf{a}^{\top} \mathcal{J}^{-1}] \begin{bmatrix} \mathbf{V} & \mathbf{I}_k \\ \mathbf{I}_k & \mathcal{J} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ -\mathcal{J}^{-1} \mathbf{a} \end{bmatrix} = \mathbf{a}^{\top} (\mathbf{V} - \mathcal{J}^{-1}) \mathbf{a}.$$

Mais ceci implique, comme  $\mathbf{a}$  est arbitraire, que la matrice  $(\mathbf{V} - \mathcal{J}^{-1})$  est semi-définie positive, ce qui correspond à ce que nous avons voulu prouver.

Ce résultat constitue un cas particulier de la **borne inférieure de Cramér-Rao**, suggérée à l'origine par Fisher (1925) dans un de ses premiers articles classiques sur l'estimation ML et énoncé sous sa forme moderne par Cramér (1946) et Rao (1945). Celle-ci est spéciale parce qu'il s'agit d'une version asymptotique du résultat d'origine. La borne inférieure de Cramér-Rao s'applique en fait à *n'importe quel* estimateur sans biais sans tenir compte de la taille de l'échantillon. Cependant, comme les estimateurs ML ne sont pas en général sans biais, seul le résultat de la version asymptotique représente un intérêt dans le contexte de l'estimation ML, et aussi avons-nous restreint notre attention au cas asymptotique.

Le fait que l'estimateur ML atteigne asymptotiquement la borne inférieure de Cramér-Rao implique que n'importe quel estimateur convergent au taux  $n^{1/2}$  peut être écrit comme la somme de l'estimateur ML et d'un autre vecteur aléatoire qui est asymptotiquement indépendant du premier. Ce résultat fournit une manière révélatrice de réfléchir à la relation entre les estimateurs efficaces et non efficaces. Pour l'établir, nous commençons par poser les définitions

$$\begin{aligned} \tilde{\mathbf{t}}^n(\boldsymbol{\theta}) &\equiv n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}), & \tilde{\mathbf{t}}(\boldsymbol{\theta}) &\equiv \text{plim}_{n \rightarrow \infty}(\tilde{\mathbf{t}}^n(\boldsymbol{\theta})), \\ \mathbf{v}^n &\equiv \hat{\mathbf{t}}^n(\boldsymbol{\theta}) - \tilde{\mathbf{t}}^n(\boldsymbol{\theta}), & \text{et } \mathbf{v} &\equiv \hat{\mathbf{t}}(\boldsymbol{\theta}) - \tilde{\mathbf{t}}(\boldsymbol{\theta}). \end{aligned} \tag{8.65}$$

Comme on peut le voir à partir des définitions (8.60) et (8.65),  $\mathbf{v}^n$  et  $\mathbf{v}$  ne dépendent pas directement de  $\boldsymbol{\theta}$ .

Nous souhaitons montrer que la matrice de covariance de  $\mathbf{v}$  et  $\tilde{\mathbf{t}}$  est une matrice égale à zéro. Cette matrice de covariance est

$$\begin{aligned} \text{Cov}_{\boldsymbol{\theta}}(\mathbf{v}, \tilde{\mathbf{t}}(\boldsymbol{\theta})) &= E_{\boldsymbol{\theta}}(\mathbf{v} \tilde{\mathbf{t}}^{\top}(\boldsymbol{\theta})) \\ &= E_{\boldsymbol{\theta}}\left((\hat{\mathbf{t}}(\boldsymbol{\theta}) - \tilde{\mathbf{t}}(\boldsymbol{\theta})) \tilde{\mathbf{t}}^{\top}(\boldsymbol{\theta})\right) \\ &= E_{\boldsymbol{\theta}}(\hat{\mathbf{t}}(\boldsymbol{\theta}) \tilde{\mathbf{t}}^{\top}(\boldsymbol{\theta})) - \mathcal{J}^{-1}(\boldsymbol{\theta}). \end{aligned} \tag{8.66}$$



En utilisant l'égalité de la matrice d'information, le résultat (8.38) peut être écrit comme

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} (\mathcal{J}(\boldsymbol{\theta}))^{-1} (n^{-1/2} \mathbf{g}(\boldsymbol{\theta})).$$

Dans la notation de (8.59) et (8.60), ceci devient

$$\tilde{\mathbf{t}}(\boldsymbol{\theta}) = \mathcal{J}^{-1}(\boldsymbol{\theta}) \mathbf{s}(\boldsymbol{\theta}).$$

Ainsi, en continuant à partir de la dernière ligne de (8.66), nous obtenons

$$\begin{aligned} \text{Cov}_{\boldsymbol{\theta}}(\mathbf{v}, \tilde{\mathbf{t}}(\boldsymbol{\theta})) &= E_{\boldsymbol{\theta}}(\hat{\mathbf{t}}(\boldsymbol{\theta}) \mathbf{s}^{\top}(\boldsymbol{\theta}) \mathcal{J}^{-1}(\boldsymbol{\theta})) - \mathcal{J}^{-1}(\boldsymbol{\theta}) \\ &= E_{\boldsymbol{\theta}}(\hat{\mathbf{t}}(\boldsymbol{\theta}) \mathbf{s}^{\top}(\boldsymbol{\theta})) \mathcal{J}^{-1}(\boldsymbol{\theta}) - \mathcal{J}^{-1}(\boldsymbol{\theta}) \\ &= \mathcal{J}^{-1}(\boldsymbol{\theta}) - \mathcal{J}^{-1}(\boldsymbol{\theta}) = \mathbf{0}. \end{aligned}$$

Le résultat fondamental (8.61) a été utilisé pour obtenir ici la dernière ligne.

Ainsi, nous concluons que

$$\hat{\mathbf{t}}(\boldsymbol{\theta}) = \tilde{\mathbf{t}}(\boldsymbol{\theta}) + \mathbf{v}, \quad (8.67)$$

où  $\mathbf{v}$  est asymptotiquement non corrélé avec  $\tilde{\mathbf{t}}$ . Si  $\hat{\mathbf{t}}$  et  $\tilde{\mathbf{t}}$  sont asymptotiquement normaux, cette corrélation asymptotiquement nulle implique par la suite une indépendance asymptotique. Une autre manière d'écrire le résultat (8.67) est

$$\hat{\boldsymbol{\theta}} \stackrel{a}{=} \tilde{\boldsymbol{\theta}} + n^{-1/2} \mathbf{v}^n.$$

Ceci montre clairement qu'un estimateur  $\hat{\boldsymbol{\theta}}$  non efficace mais convergent peut toujours être décomposé, asymptotiquement, en la somme d'un estimateur ML  $\tilde{\boldsymbol{\theta}}$  asymptotiquement efficace et d'une autre variable aléatoire, qui tend vers zéro quand  $n \rightarrow \infty$  et est asymptotiquement non corrélée avec l'estimateur efficace. Evidemment, tout l'éventail des estimateurs asymptotiquement normaux et convergents peut être généré à partir de l'estimateur ML  $\tilde{\boldsymbol{\theta}}$  en lui additionnant des variables aléatoires multivariées normales d'espérances nulles indépendantes de  $\tilde{\boldsymbol{\theta}}$ . On peut imaginer que celles-ci soient des bruits parasitant le signal efficace émis par  $\tilde{\boldsymbol{\theta}}$ . L'interprétation du résultat de Cramér-Rao est assez évidente à présent: comme la variance de la somme de deux variables aléatoires indépendantes est la somme de leurs variances respectives, la matrice semi-définie positive qui correspond à la différence entre les matrices de covariance de  $\hat{\boldsymbol{\theta}}$  et  $\tilde{\boldsymbol{\theta}}$  est précisément la matrice de covariance (peut-être dégénérée) du vecteur des variables de bruit  $n^{-1/2} \mathbf{v}$ .

Ces résultats pour les estimateurs ML sont similaires, mais beaucoup plus forts que les résultats obtenus pour les moindres carrés non linéaires dans la Section 5.5. Nous y avons vu que n'importe quel estimateur convergent mais non efficace qui est asymptotiquement linéaire pour les aléas peut être écrit comme la somme de l'estimateur efficace et d'une variable aléatoire (ou vecteur) qui est asymptotiquement non corrélée avec l'estimateur efficace. La démonstration du Théorème de Gauss-Markov fournissait également un résultat similaire.

## 8.9 LES TROIS STATISTIQUES DE TEST CLASSIQUES

Une des caractéristiques attrayantes de l'estimation ML est que les statistiques de test basées sur les trois principes dont nous avons discuté pour la première fois dans le Chapitre 3 —le principe du rapport de vraisemblance, le principe du multiplicateur de Lagrange et le principe de Wald —sont toujours disponibles et sont souvent faciles à calculer. Ces trois principes de test d'hypothèse furent énoncés pour la première fois dans le contexte de l'estimation ML, et certains auteurs utilisent encore les termes de “rapport de vraisemblance”, “multiplicateur de Lagrange”, et “Wald” dans le seul contexte des tests basés sur les estimations ML. Dans cette section, nous fournissons une introduction à ce que l'on désigne souvent sous le nom des **trois tests classiques**. Ces trois statistiques de test possèdent la même distribution asymptotique sous l'hypothèse nulle; s'il y a  $r$  contraintes d'égalité, elles sont distribuées suivant une distribution du  $\chi^2(r)$ . En effet, elles tendent réellement vers la même variable aléatoire asymptotiquement, à la fois sous l'hypothèse nulle et sous la série des DGP qui sont proches de l'hypothèse nulle dans un certains sens. Un traitement approprié de ces résultats importants nécessite plus de développements que nous n'en disposons dans cette section. Ainsi, nous remettons celui-ci au Chapitre 13, qui fournit une discussion beaucoup plus détaillée des trois statistiques de test classiques.

Conceptuellement, le plus simple des trois tests classiques est le **rapport de vraisemblance**, ou test **LR**. La statistique de test est simplement deux fois la différence entre les valeurs contrainte et non contrainte de la fonction de logvraisemblance,

$$2(\ell(\hat{\theta}) - \ell(\tilde{\theta})), \quad (8.68)$$

où  $\hat{\theta}$  désigne l'estimation ML non contrainte de  $\theta$ ,  $\tilde{\theta}$  désigne l'estimation ML soumise aux  $r$  contraintes distinctes, et où la dépendance de  $\ell$  à  $\mathbf{y}$  a été supprimée pour simplifier la notation. Le nom de la statistique LR provient du fait que (8.68) est égale à

$$2 \log \left( \frac{L(\hat{\theta})}{L(\tilde{\theta})} \right),$$

ou deux fois le logarithme du rapport des fonctions de vraisemblance. Elle est très facile à calculer lorsqu'à la fois les estimations contraintes et les non contraintes sont disponibles, et c'est une de ses caractéristiques les plus attrayantes.

Pour dériver la distribution asymptotique de la statistique LR, il faut calculer un développement en série de Taylor au second ordre de  $\ell(\tilde{\theta})$  autour de  $\hat{\theta}$ . Bien que nous ne terminerons pas la construction de cette statistique dans cette section, il est révélateur de parcourir les premières étapes. Le résultat du développement en série de Taylor est

$$\ell(\tilde{\theta}) \cong \ell(\hat{\theta}) + \frac{1}{2}(\tilde{\theta} - \hat{\theta})^\top \mathbf{H}(\hat{\theta})(\tilde{\theta} - \hat{\theta}). \quad (8.69)$$

Ici, il n'y a pas de terme du premier ordre parce que  $\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$  grâce aux conditions du premier ordre (8.12). En ordonnant les termes de (8.69) nous obtenons

$$\begin{aligned} 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})) &\cong -(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \mathbf{H}(\hat{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\ &\stackrel{a}{=} (n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}))^\top \mathbf{J}(\hat{\boldsymbol{\theta}})(n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})). \end{aligned} \quad (8.70)$$

Cet exercice permet d'expliquer la provenance du facteur de 2 dans la définition de la statistique LR. La prochaine étape consisterait à remplacer  $n^{1/2}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$  dans (8.70) par

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

et d'utiliser ensuite le résultat (8.38), simultanément avec un résultat analogue pour les estimations contraintes que nous obtiendrons sous peu, pour établir la distribution asymptotique de la statistique LR. Nous réaliserons ceci dans le Chapitre 13.

Nous portons maintenant notre attention sur le **multiplicateur de Lagrange**, ou test **LM**. En effet, cette statistique de test porte deux noms et prend deux formes différentes, qui s'avèrent être numériquement identiques si la même estimation de la matrice d'information est utilisée pour les calculer. Une forme, proposée à l'origine par Rao (1948), est appelée la **forme score du test LM**, ou simplement le **test score**, et est calculée en utilisant le gradient ou le vecteur score du modèle non contraint évalué avec les estimations contraintes. L'autre forme, qui donne au test son nom, a été proposée par Aitchison et Silvey (1958, 1960) et Silvey (1959). Cette dernière forme est calculée en utilisant le vecteur des multiplicateurs de Lagrange qui émerge si on maximise la fonction de vraisemblance soumise aux contraintes au moyen d'un Lagrangien. Les économètres utilisent généralement le test LM sous sa forme score mais insistent néanmoins pour le nommer test LM, peut-être parce que les multiplicateurs de Lagrange sont aussi largement utilisés en économétrie. Les références sur les tests LM en économétrie sont Breusch et Pagan (1980) et Engle (1982a, 1984). Buse (1982) fournit une discussion intuitive des relations entre les tests LR, LM, et Wald.

Une manière de maximiser  $\ell(\boldsymbol{\theta})$  soumise aux contraintes exactes

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}, \quad (8.71)$$

où  $\mathbf{r}(\boldsymbol{\theta})$  est un vecteur de dimension  $r$  avec  $r \leq k$ , consiste à maximiser simultanément le Lagrangien

$$\ell(\boldsymbol{\theta}) - \mathbf{r}^\top(\boldsymbol{\theta})\boldsymbol{\lambda}$$

par rapport à  $\boldsymbol{\theta}$  et à le minimiser par rapport au vecteur de dimension  $r$   $\boldsymbol{\lambda}$  des multiplicateurs de Lagrange. Les conditions du premier ordre qui caractérisent la solution de ce problème sont

$$\begin{aligned} \mathbf{g}(\tilde{\boldsymbol{\theta}}) - \mathbf{R}^\top(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ \mathbf{r}(\tilde{\boldsymbol{\theta}}) &= \mathbf{0}, \end{aligned} \quad (8.72)$$

où  $\mathbf{R}(\boldsymbol{\theta})$  est une matrice de dimension  $r \times k$  avec comme élément type  $\partial r_i(\boldsymbol{\theta})/\partial \theta_j$ .

Nous sommes intéressés par la distribution de  $\tilde{\boldsymbol{\lambda}}$  sous l'hypothèse nulle, aussi supposons-nous que le DGP satisfait (8.71) avec le vecteur paramétrique  $\boldsymbol{\theta}_0$ . La valeur du vecteur  $\boldsymbol{\lambda}$  des multiplicateurs de Lagrange si  $\tilde{\boldsymbol{\theta}}$  était égal à  $\boldsymbol{\theta}_0$  devrait être égale à zéro. Ainsi, il semble naturel de prendre un développement en série de Taylor au premier ordre des conditions du premier ordre (8.72) autour du point  $(\boldsymbol{\theta}_0, \mathbf{0})$ . Ceci donne

$$\begin{aligned} \mathbf{g}(\boldsymbol{\theta}_0) + \mathbf{H}(\bar{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \mathbf{R}^\top(\bar{\boldsymbol{\theta}})\tilde{\boldsymbol{\lambda}} &= \mathbf{0} \\ -\mathbf{R}(\ddot{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= \mathbf{0}, \end{aligned}$$

où  $\bar{\boldsymbol{\theta}}$  et  $\ddot{\boldsymbol{\theta}}$  désignent les valeurs de  $\boldsymbol{\theta}$  qui se situent entre  $\tilde{\boldsymbol{\theta}}$  et  $\boldsymbol{\theta}_0$ . Ces équations peuvent être réécrites comme

$$\begin{bmatrix} -\mathbf{H}(\bar{\boldsymbol{\theta}}) & \mathbf{R}^\top(\bar{\boldsymbol{\theta}}) \\ \mathbf{R}(\ddot{\boldsymbol{\theta}}) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \tilde{\boldsymbol{\lambda}} \end{bmatrix} = \begin{bmatrix} \mathbf{g}(\boldsymbol{\theta}_0) \\ \mathbf{0} \end{bmatrix}. \quad (8.73)$$

Si nous multiplions  $\mathbf{H}(\bar{\boldsymbol{\theta}})$  par  $n^{-1}$ ,  $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$  par  $n^{1/2}$ ,  $\mathbf{g}(\boldsymbol{\theta}_0)$  par  $n^{-1/2}$ , et  $\tilde{\boldsymbol{\lambda}}$  par  $n^{-1/2}$ , nous ne changeons pas l'égalité dans (8.73), et nous transformons toutes les quantités qui y apparaissent en des quantités  $O(1)$ . Les lecteurs peuvent vouloir vérifier que ces facteurs de  $n$  sont en effet les plus appropriés et, en particulier, que  $\tilde{\boldsymbol{\lambda}}$  doit être multiplié par  $n^{-1/2}$ . En utilisant le fait que  $\tilde{\boldsymbol{\theta}}$  et par conséquent  $\bar{\boldsymbol{\theta}}$  et  $\ddot{\boldsymbol{\theta}}$  sont convergents, en appliquant une loi des grands nombres convenable à  $n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}})$ , et en résolvant les équations du système résultant, nous obtenons

$$\begin{bmatrix} n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ n^{-1/2}\tilde{\boldsymbol{\lambda}} \end{bmatrix} \stackrel{a}{=} \begin{bmatrix} -\mathcal{H}_0 & \mathbf{R}_0^\top \\ \mathbf{R}_0 & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0) \\ \mathbf{0} \end{bmatrix}, \quad (8.74)$$

où  $\mathcal{H}_0$  désigne  $\mathcal{H}(\boldsymbol{\theta}_0)$  et  $\mathbf{R}_0$  désigne  $\mathbf{R}(\boldsymbol{\theta}_0)$ .

Le système des équations (8.74) est, pour le cas contraint, l'équivalent de l'équation (8.38) pour le cas non contraint. La première chose à noter, le concernant, est que les  $k$  éléments de  $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  et les  $r$  éléments de  $n^{-1/2}\tilde{\boldsymbol{\lambda}}$  dépendent tous du vecteur de dimension  $k$  aléatoire  $n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0)$ . Nous avons déjà vu que, sous des conditions de régularité standards, ce dernier est asymptotiquement normalement distribué avec un vecteur d'espérances nulles et une matrice de covariance  $\mathcal{J}(\boldsymbol{\theta}_0)$ . Ainsi à partir de (8.74) nous voyons qu'à la fois  $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  et  $n^{-1/2}\tilde{\boldsymbol{\lambda}}$  doivent être asymptotiquement normalement distribués. Observons que le vecteur de dimension  $(k+r)$  dans le membre de gauche de (8.74) doit avoir une matrice de covariance singulière, car son rang ne peut pas excéder  $k$ , qui est le rang de  $\mathcal{J}(\boldsymbol{\theta}_0)$ .

En inversant analytiquement la matrice partitionnée et en multipliant ensuite les deux facteurs du membre de droite de (8.74), il est possible d'obtenir

assez facilement, bien que cela soit quelque peu ennuyeux, les expressions de  $n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  et de  $n^{-1/2}\tilde{\boldsymbol{\lambda}}$ . Celles-ci sont

$$n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\mathcal{H}_0^{-1}(\mathbf{I} - \mathbf{R}_0^\top(\mathbf{R}_0\mathcal{H}_0^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0\mathcal{H}_0^{-1})(n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0))$$

et

$$n^{-1/2}\tilde{\boldsymbol{\lambda}} \stackrel{a}{=} (\mathbf{R}_0\mathcal{H}_0^{-1}\mathbf{R}_0^\top)^{-1}\mathbf{R}_0\mathcal{H}_0^{-1}(n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0)).$$

À partir de la seconde de ces expressions, de la normalité asymptotique de  $n^{-1/2}\mathbf{g}(\boldsymbol{\theta}_0)$ , et de l'égalité de la matrice d'information, il est facile de voir que

$$n^{-1/2}\tilde{\boldsymbol{\lambda}} \stackrel{a}{\sim} N(\mathbf{0}, (\mathbf{R}_0\mathcal{J}_0^{-1}\mathbf{R}_0^\top)^{-1}). \quad (8.75)$$

Maintenant, il est simple de dériver le test du multiplicateur de Lagrange sous sa forme LM. La statistique de test est simplement une forme quadratique du vecteur de dimension  $r$   $n^{-1/2}\tilde{\boldsymbol{\lambda}}$ :

$$(n^{-1/2}\tilde{\boldsymbol{\lambda}})^\top(\tilde{\mathbf{R}}\tilde{\mathcal{J}}^{-1}\tilde{\mathbf{R}}^\top)(n^{-1/2}\tilde{\boldsymbol{\lambda}}) = \frac{1}{n}\tilde{\boldsymbol{\lambda}}^\top\tilde{\mathbf{R}}\tilde{\mathcal{J}}^{-1}\tilde{\mathbf{R}}^\top\tilde{\boldsymbol{\lambda}}. \quad (8.76)$$

Ici,  $\tilde{\mathcal{J}}$  peut être n'importe quelle matrice qui utilise les estimations contraintes  $\tilde{\boldsymbol{\theta}}$  pour estimer  $\mathcal{J}(\boldsymbol{\theta}_0)$  de manière convergente. Différentes variantes de la statistique LM utiliseront différentes estimations de  $\mathcal{J}(\boldsymbol{\theta}_0)$ . Il est évident à partir de (8.75), que sous les conditions de régularité standards cette statistique de test sera asymptotiquement distribuée suivant une  $\chi^2(r)$  sous l'hypothèse nulle.

La statistique LM (8.76) est numériquement égale à un test basé sur le vecteur score  $\mathbf{g}(\tilde{\boldsymbol{\theta}})$ . Du premier ensemble des conditions du premier ordre (8.72),  $\mathbf{g}(\tilde{\boldsymbol{\theta}}) = \mathbf{R}^\top\tilde{\boldsymbol{\lambda}}$ . Si l'on substitue  $\mathbf{g}(\tilde{\boldsymbol{\theta}})$  à  $\mathbf{R}^\top\tilde{\boldsymbol{\lambda}}$  dans (8.76) nous aboutissons à la forme score du test LM,

$$\frac{1}{n}\tilde{\mathbf{g}}^\top\tilde{\mathcal{J}}^{-1}\tilde{\mathbf{g}}. \quad (8.77)$$

Dans la pratique, cette forme score est souvent plus utile que la forme LM parce que, comme les estimations contraintes sont rarement obtenues via un Lagrangien,  $\tilde{\mathbf{g}}$  est généralement facilement disponible alors que typiquement  $\tilde{\boldsymbol{\lambda}}$  ne l'est pas. Cependant, la construction du test via les multiplicateurs de Lagrange est révélatrice, car elle montre clairement la provenance des  $r$  degrés de liberté.

Le troisième des trois tests classiques est le **test de Wald**. Ce test est très facile à dériver. Il consiste à savoir si le vecteur des contraintes, évaluées à l'aide des estimations non contraintes est suffisamment proche du vecteur nul pour que les contraintes soient plausibles. Dans le cas des contraintes (8.71), le test de Wald est basé sur le vecteur  $\mathbf{r}(\hat{\boldsymbol{\theta}})$ , qui devrait tendre asymptotiquement vers un vecteur nul si les contraintes sont valables. Comme nous l'avons vu dans les Sections 8.5 et 8.6,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{\sim} N(\mathbf{0}, \mathcal{J}^{-1}(\boldsymbol{\theta}_0)).$$

Un développement en série de Taylor de  $\mathbf{r}(\hat{\boldsymbol{\theta}})$  autour de  $\boldsymbol{\theta}_0$  donne  $\mathbf{r}(\hat{\boldsymbol{\theta}}) \cong \mathbf{R}_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ . Ainsi,

$$\mathbf{V}(n^{1/2}\mathbf{r}(\hat{\boldsymbol{\theta}})) \stackrel{a}{=} \mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top.$$

Il s'ensuit qu'une statistique de test appropriée est

$$n\mathbf{r}^\top(\hat{\boldsymbol{\theta}})(\hat{\mathbf{R}}\hat{\mathcal{J}}^{-1}\hat{\mathbf{R}}^\top)^{-1}\mathbf{r}(\hat{\boldsymbol{\theta}}), \quad (8.78)$$

où  $\hat{\mathcal{J}}$  désigne n'importe quelle estimation de  $\mathcal{J}(\boldsymbol{\theta}_0)$  basée sur les estimations non contraintes  $\hat{\boldsymbol{\theta}}$ . Différentes variantes du test de Wald utiliseront différentes estimations de  $\mathcal{J}(\boldsymbol{\theta}_0)$ . Il est facile de voir qu'étant données les conditions de régularité adéquates, la statistique de test (8.78) sera asymptotiquement distribuée suivant une  $\chi^2(r)$  sous l'hypothèse nulle.

La propriété fondamentale des trois statistiques des test classiques est que sous l'hypothèse nulle, quand  $n \rightarrow \infty$ , elles tendent toutes vers la même variable aléatoire, qui est distribuée suivant une  $\chi^2(r)$ . Nous prouverons ce résultat au cours du Chapitre 13. La conséquence est que, avec de grands échantillons, le choix parmi les trois importe peu. Si à la fois  $\hat{\boldsymbol{\theta}}$  et  $\tilde{\boldsymbol{\theta}}$  sont faciles à calculer, il est intéressant d'utiliser le test LR. Si  $\tilde{\boldsymbol{\theta}}$  est facile à calculer mais que  $\hat{\boldsymbol{\theta}}$  ne l'est pas, comme cela est souvent le cas pour les tests de spécification de modèle, alors le test LM devient attrayant. Si d'un autre côté  $\hat{\boldsymbol{\theta}}$  est facile à calculer mais  $\tilde{\boldsymbol{\theta}}$  ne l'est pas, comme cela peut être le cas quand nous sommes intéressés par les contraintes non linéaires imposées à un modèle linéaire, alors le test de Wald devient attrayant. Quand la taille de l'échantillon n'est pas grande, un choix pertinent parmi les trois tests est compliqué par le fait qu'ils peuvent avoir des propriétés avec des échantillons finis très différentes, qui peuvent par la suite différer formidablement selon les variantes alternatives des tests LM et Wald. Ceci rend le choix des tests plutôt plus compliqué en pratique que ce que la théorie asymptotique ne le suggère.

## 8.10 LES MODÈLES DE RÉGRESSION NON LINÉAIRE

Dans cette section, nous discutons des possibilités de l'usage de la méthode du maximum de vraisemblance pour l'estimation des modèles de régression univarié non linéaire. Quand les aléas sont supposés être normalement et indépendamment distribués avec une variance constante, l'estimation ML de ces modèles est, du moins en ce qui concerne l'estimation des paramètres de la fonction de régression, numériquement identique à l'estimation NLS. L'exercice présente néanmoins un intérêt. Tout d'abord, il fournit une illustration concrète de la manière d'utiliser la méthode du maximum de vraisemblance. Deuxièmement, il fournit une matrice de covariance asymptotique pour les estimations de  $\boldsymbol{\beta}$  et  $\sigma$  conjointement, alors que les NLS ne la calculent que pour les estimations de  $\boldsymbol{\beta}$ . Finalement, en considérant certaines extensions du modèle de régression normal, nous sommes capables de démontrer la puissance de l'estimation ML.

La classe des modèles que nous considérerons est

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (8.79)$$

où la fonction de régression  $\mathbf{x}(\boldsymbol{\beta})$  satisfait les conditions pour les Théorèmes 5.1 et 5.2, et les données sont supposées avoir été générées par un cas particulier de (8.79). Le vecteur paramétrique  $\boldsymbol{\beta}$  est supposé être de longueur  $k$ , ce qui implique qu'il y a  $k + 1$  paramètres à estimer. La notation " $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ " signifie que le vecteur des aléas  $\mathbf{u}$  est supposé être distribué suivant une loi normale multivariée de vecteur d'espérance zéro et de matrice de covariance  $\sigma^2 \mathbf{I}$ . Ainsi, les aléas individuels  $u_t$  sont indépendants, chacun étant distribué suivant la  $N(0, \sigma^2)$ . La fonction de densité de  $u_t$  est

$$f(u_t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{u_t^2}{2\sigma^2}\right).$$

Afin de construire la fonction de vraisemblance, nous avons besoin de la fonction de densité de  $y_t$  plutôt que de celle de  $u_t$ . Ceci nous demande d'utiliser un résultat standard en statistique qui est établi dans l'Annexe B.

Le résultat en question indique que si une variable aléatoire  $x_1$  a une fonction de densité  $f_1(x_1)$  et si une autre variable aléatoire  $x_2$  lui est reliée par

$$x_1 = h(x_2),$$

où la fonction  $h(\cdot)$  est monotone et continûment différentiable, alors la fonction de densité de  $x_2$  est donnée par

$$f_2(x_2) = f_1(h(x_2)) \left| \frac{\partial h(x_2)}{\partial x_2} \right|.$$

Ici, le second facteur est la valeur absolue du Jacobien de la transformation. Dans de nombreux cas, comme nous le verrons plus tard, sa présence fait apparaître les **termes Jacobiens** dans les fonctions de logvraisemblance. Cependant, dans ce cas, la fonction qui relie  $u_t$  à  $y_t$  est

$$u_t = y_t - x_t(\boldsymbol{\beta}).$$

Le facteur Jacobien  $|\partial u_t / \partial y_t|$  est alors égal à l'unité. Ainsi, nous concluons que la fonction de densité de  $y_t$  est

$$\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{(y_t - x_t(\boldsymbol{\beta}))^2}{2\sigma^2}\right). \quad (8.80)$$

La contribution à la fonction de logvraisemblance apportée par la  $t^{\text{ième}}$  observation est le logarithme de (8.80),

$$\ell_t(y_t, \boldsymbol{\beta}, \sigma) = -\frac{1}{2} \log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2} (y_t - x_t(\boldsymbol{\beta}))^2.$$

Comme toutes les informations sont indépendantes, la fonction de logvraisemblance elle-même correspond précisément à la somme des contributions  $\ell_t(y_t, \beta, \sigma)$  sur tout  $t$ , ou

$$\begin{aligned}\ell(\mathbf{y}, \beta, \sigma) &= -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^n (y_t - x_t(\beta))^2 \\ &= -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x}(\beta))^\top (\mathbf{y} - \mathbf{x}(\beta)).\end{aligned}\quad (8.81)$$

La première étape dans la maximisation de  $\ell(\mathbf{y}, \beta, \sigma)$  consiste à la concentrer par rapport à  $\sigma$ , comme cela fut expliqué dans la Section 8.7. La différentiation de la seconde ligne de (8.81) par rapport à  $\sigma$  et l'égalisation de la dérivée à zéro donnent

$$\frac{\partial \ell(\mathbf{y}, \beta, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} (\mathbf{y} - \mathbf{x}(\beta))^\top (\mathbf{y} - \mathbf{x}(\beta)) = 0,$$

et la résolution de cette équation produit le résultat

$$\hat{\sigma}(\beta) = \left( \frac{1}{n} (\mathbf{y} - \mathbf{x}(\beta))^\top (\mathbf{y} - \mathbf{x}(\beta)) \right)^{1/2}.$$

Ici la notation  $\hat{\sigma}(\beta)$  signifie que l'estimation ML de  $\sigma$  est maintenant une fonction de  $\beta$ . Notons que nous avons divisé par  $n$  plutôt que par  $n - k$ . Si nous pouvions évaluer  $\hat{\sigma}^2(\beta)$  à la véritable valeur  $\beta_0$ , nous obtiendrions une estimation non biaisée de  $\sigma^2$ . Cependant, nous l'évaluons en fait à l'estimation ML  $\hat{\beta}$ , qui, comme nous le voyons, est égale à l'estimation NLS. Ainsi, comme nous l'avons vu dans la Section 3.2,  $\hat{\sigma}^2$  doit être biaisée vers le bas en tant qu'estimateur de  $\sigma^2$ .

La substitution de  $\hat{\sigma}(\beta)$  dans la seconde ligne de (8.81) permet de construire la fonction de logvraisemblance concentrée

$$\begin{aligned}\ell^c(\mathbf{y}, \beta) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log\left(\frac{1}{n} (\mathbf{y} - \mathbf{x}(\beta))^\top (\mathbf{y} - \mathbf{x}(\beta))\right) - \frac{n}{2} \\ &= C - \frac{n}{2} \log\left((\mathbf{y} - \mathbf{x}(\beta))^\top (\mathbf{y} - \mathbf{x}(\beta))\right),\end{aligned}\quad (8.82)$$

où  $C$  est un terme constant. Le second terme dans (8.82) est moins  $n/2$  fois le logarithme de la somme des résidus au carré. Ainsi, nous voyons que *maximiser* la fonction de logvraisemblance concentrée est équivalent à *minimiser*  $SSR(\beta)$ . Les estimations ML  $\hat{\beta}$  seront simplement les estimations NLS avec lesquelles nous sommes déjà familiers.

Le terme constant dans (8.82) est en fait

$$\frac{n}{2} (\log(n) - 1 - \log(2\pi)).$$



Comme cette expression ne dépend pas de  $\beta$ , elle peut être ignorée dans toutes les utilisations sauf en fait pour le calcul de la valeur de  $\ell(\mathbf{y}, \beta, \sigma)$ . De telles constantes sont souvent complètement ignorées dans un travail théorique et sont même parfois ignorées par des programmes informatiques, et le résultat de tout ceci est que les valeurs des fonctions de logvraisemblance pour un même modèle et un même ensemble de données reportées par différents programmes peuvent parfois différer.

Le fait que l'estimateur ML  $\hat{\beta}$  pour la classe des modèles (8.79) corresponde exactement à l'estimateur NLS comporte une importante implication. Comme nous l'avons vu dans la Section 8.8, les estimateurs ML sont asymptotiquement efficaces. Ainsi, l'estimateur NLS sera asymptotiquement efficace à chaque fois que les aléas sont normalement et indépendamment distribués avec une variance constante. Cependant, si les aléas ont une quelqu'autre distribution connue, l'estimateur ML diffèrera en général de celui des NLS et sera plus efficace que ce dernier (voir plus loin pour un exemple extrême). Ainsi, bien que l'estimateur NLS soit convergent sous de très faibles conditions sur la distribution des aléas, comme nous l'avons vu dans la Section 5.3, et soit efficace dans la classe des estimateurs asymptotiquement linéaires qui sont applicables sous ces conditions peu restrictives, il ne coïncide avec l'estimateur ML efficace que si les aléas sont supposés être normalement distribués. La signification de tout ceci est la suivante. Si la seule hypothèse que l'on veut formuler concernant les aléas est qu'ils satisfassent les conditions de régularité pour les NLS, alors l'estimateur NLS est asymptotiquement efficace dans la classe des estimateurs asymptotiquement linéaires et convergents des paramètres de la fonction de régression. Cependant, si l'on est prêt à fournir l'effort de spécifier la véritable distribution des aléas, alors l'estimateur ML sera en général plus efficace, à condition que la spécification présumée des aléas soit correcte. L'estimateur ML ne sera pas plus efficace dans le cas où les aléas sont supposés être normaux, puisqu'alors les estimateurs ML et NLS seront équivalents.

Dans la Section 8.6, nous avons vu que si  $\hat{\theta}$  est un vecteur d'estimations ML, alors le vecteur  $n^{1/2}(\hat{\theta} - \theta_0)$  est asymptotiquement normalement distribué avec un vecteur d'espérance zéro et une matrice de covariance égale à l'inverse de la matrice d'information asymptotique  $\mathcal{J}(\theta_0)$ . Ce résultat signifie qu'il est presque toujours intéressant de calculer  $\mathcal{J}(\theta)$  pour n'importe quel modèle qui est estimé par maximum de vraisemblance. Nous avons vu qu'il y a en général deux manières de procéder. L'une consiste à trouver l'opposée de la limite en probabilité de  $n^{-1}$  fois la matrice Hessienne, et l'autre consiste à trouver la limite en probabilité de  $n^{-1}$  fois  $\mathbf{G}^\top(\theta)\mathbf{G}(\theta)$ , où  $\mathbf{G}(\theta)$  est la matrice CG. Ces deux méthodes entraîneront la même réponse, s'il est tout à fait faisable de calculer  $\mathcal{J}(\theta)$ , bien qu'une approche puisse être plus facile que l'autre dans certaines situations données.

Pour le modèle de régression non linéaire (8.79), le vecteur paramétrique  $\theta$  est le vecteur  $[\beta : \sigma]$ . Nous calculons à présent la matrice d'information

asymptotique  $\mathcal{J}(\boldsymbol{\beta}, \sigma)$  pour ce modèle en utilisant la seconde méthode, basée sur la matrice CG, qui ne nécessite que les dérivées premières. Il s'agit d'un bon exercice que de répéter la construction en utilisant la matrice Hessienne, qui nécessite les dérivées secondes, et de vérifier que cela produit les mêmes résultats. La dérivée première de  $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$  par rapport à  $\beta_i$  est

$$\frac{\partial \ell_t}{\partial \beta_i} = \frac{1}{\sigma^2} (y_t - x_t(\boldsymbol{\beta})) X_{ti}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} e_t(\boldsymbol{\beta}) X_{ti}(\boldsymbol{\beta}), \quad (8.83)$$

où  $e_t(\boldsymbol{\beta}) \equiv y_t - x_t(\boldsymbol{\beta})$  et, comme d'habitude,  $X_{ti}(\boldsymbol{\beta}) \equiv \partial x_t(\boldsymbol{\beta}) / \partial \beta_i$ . La dérivée première de  $\ell_t(y_t, \boldsymbol{\beta}, \sigma)$  par rapport à  $\sigma$  est

$$\frac{\partial \ell_t}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(y_t - x_t(\boldsymbol{\beta}))^2}{\sigma^3} = -\frac{1}{\sigma} + \frac{e_t^2(\boldsymbol{\beta})}{\sigma^3}. \quad (8.84)$$

Les expressions (8.83) et (8.84) sont tout ce dont nous avons besoin pour calculer la matrice d'information en utilisant la matrice CG. La colonne de cette matrice qui correspond à  $\sigma$  aura l'élément type (8.84), tandis que les  $k$  colonnes restantes, qui correspondent aux  $\beta_i$ , auront l'élément type (8.83).

L'élément de  $\mathcal{J}(\boldsymbol{\beta}, \sigma)$  correspondant à  $\beta_i$  et  $\beta_j$  est

$$\mathcal{J}(\beta_i, \beta_j) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \frac{e_t^2(\boldsymbol{\beta})}{\sigma^4} X_{ti}(\boldsymbol{\beta}) X_{tj}(\boldsymbol{\beta}) \right).$$

Comme  $e_t^2(\boldsymbol{\beta})$  a une espérance de  $\sigma^2$  sous le DGP caractérisé par  $(\boldsymbol{\beta}, \sigma)$  et est indépendant de  $\mathbf{X}(\boldsymbol{\beta})$ , nous pouvons le remplacer ici par  $\sigma^2$  pour obtenir

$$\mathcal{J}(\beta_i, \beta_j) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \frac{1}{\sigma^2} X_{ti}(\boldsymbol{\beta}) X_{tj}(\boldsymbol{\beta}) \right).$$

Ainsi, nous voyons que le bloc entier  $(\boldsymbol{\beta}, \boldsymbol{\beta})$  de la matrice d'information asymptotique est

$$\frac{1}{\sigma^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top(\boldsymbol{\beta}) \mathbf{X}(\boldsymbol{\beta}) \right). \quad (8.85)$$

L'élément de  $\mathcal{J}(\boldsymbol{\beta}, \sigma)$  correspondant à  $\sigma$  est

$$\begin{aligned} \mathcal{J}(\sigma, \sigma) &= \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \left( \frac{1}{\sigma^2} + \frac{e_t^4(\boldsymbol{\beta})}{\sigma^6} - \frac{2e_t^2(\boldsymbol{\beta})}{\sigma^4} \right) \right) \\ &= \frac{1}{n} \left( \frac{n}{\sigma^2} + \frac{3n\sigma^4}{\sigma^6} - \frac{2n\sigma^2}{\sigma^4} \right) \\ &= \frac{2}{\sigma^2}. \end{aligned} \quad (8.86)$$

Ici, nous avons utilisé les faits que, sous le DGP caractérisé par  $(\boldsymbol{\beta}, \sigma)$ ,  $E(e_t^2(\boldsymbol{\beta})) = \sigma^2$  et  $E(e_t^4(\boldsymbol{\beta})) = 3\sigma^4$ , la dernière égalité étant une propriété bien connue de la distribution normale (consulter la Section 2.6 et l'Annexe B).

Finalement, l'élément de  $\mathcal{J}(\boldsymbol{\beta}, \sigma)$  correspondant à  $\beta_i$  et  $\sigma$  est

$$\begin{aligned} \mathcal{J}(\beta_i, \sigma) &= \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \left( -\frac{e_t(\boldsymbol{\beta}) X_{ti}(\boldsymbol{\beta})}{\sigma^3} + \frac{e_t^3(\boldsymbol{\beta}) X_{ti}(\boldsymbol{\beta})}{\sigma^5} \right) \right) \\ &= 0. \end{aligned} \quad (8.87)$$

Les éléments sont nuls parce que, sous le DGP caractérisé par  $(\boldsymbol{\beta}, \sigma)$ ,  $e_t(\boldsymbol{\beta})$  est indépendant de  $\mathbf{X}(\boldsymbol{\beta})$ , et le fait que les aléas soient normalement distribués implique que  $E(e_t(\boldsymbol{\beta})) = E(e_t^3(\boldsymbol{\beta})) = 0$ .

En collectant les résultats (8.85), (8.86), et (8.87), nous concluons que

$$\mathcal{J}(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2} \begin{bmatrix} \text{plim}(n^{-1} \mathbf{X}^\top(\boldsymbol{\beta}) \mathbf{X}(\boldsymbol{\beta})) & \mathbf{0} \\ \mathbf{0}^\top & 2 \end{bmatrix}. \quad (8.88)$$

Nos résultats sur la distribution asymptotique des estimateurs ML (Sections 8.5 et 8.6) nous permettent de conclure que

$$\begin{bmatrix} n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ n^{1/2}(\hat{\sigma} - \sigma_0) \end{bmatrix} \underset{a}{\sim} N \left( \mathbf{0}, \begin{bmatrix} \sigma_0^2 \text{plim}(n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} & \mathbf{0} \\ \mathbf{0}^\top & \sigma_0^2/2 \end{bmatrix} \right), \quad (8.89)$$

où  $\boldsymbol{\beta}_0$  et  $\sigma_0$  désignent les valeurs de  $\boldsymbol{\beta}$  et  $\sigma$  sous le DGP, et  $\mathbf{X}_0$  désigne  $\mathbf{X}(\boldsymbol{\beta}_0)$ . Parce que la matrice d'information (8.88) est bloc-diagonale entre le bloc  $(\boldsymbol{\beta}, \boldsymbol{\beta})$  et le bloc  $(\sigma, \sigma)$  (qui est un scalaire), son inverse est simplement la matrice qui se compose de chaque bloc inversé séparément. Comme nous le verrons dans le Chapitre 9, ce type de bloc-diagonalité est une propriété très importante des modèles de régression avec erreurs normales.

A partir de (8.89), nous voyons que la matrice de covariance de  $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$  est la même matrice de covariance asymptotique préalablement établie pour les estimations NLS des paramètres d'une fonction de régression, ce qui n'est pas surprenant car  $\hat{\boldsymbol{\beta}}$  est simplement un vecteur d'estimations NLS. Mais ici nous l'avons dérivée comme un cas particulier des résultats généraux de la Section 8.6 sur la distribution asymptotique des estimateurs ML. Le résultat selon lequel la variance asymptotique de  $n^{1/2}(\hat{\sigma} - \sigma_0)$  est  $\sigma_0^2/2$  est nouveau. Comme nous l'avons vu dans le Chapitre 5, la méthode des moindres carrés non linéaires ne produit pas directement une estimation de  $\sigma$  bien qu'il soit facile d'en construire plusieurs estimations, une fois que le vecteur  $\hat{\boldsymbol{\beta}}$  a été obtenu. La méthode du maximum de vraisemblance, couplée avec l'hypothèse de normalité, produit directement une estimation de  $\sigma$  et aussi une mesure de la variabilité de cette estimation. Cependant, cette dernière n'est en général

valide que sous l'hypothèse de normalité. De plus, comme nous en avons discuté plus tôt, l'estimation ML  $\hat{\sigma}^2 = n^{-1}SSR(\hat{\beta})$  est biaisée vers le bas, et en pratique il peut alors être préférable d'utiliser  $s^2 = (n - k)^{-1}SSR(\hat{\beta})$ .

Dans la dérivation de (8.88) et (8.89), nous avons choisi d'écrire la matrice d'information en termes de  $\beta$  et de  $\sigma$ . De nombreux auteurs choisissent de l'écrire en termes de  $\beta$  et de  $\sigma^2$ . Le résultat équivalent à (8.89) dans cette paramétrisation alternative est

$$\begin{bmatrix} n^{1/2}(\hat{\beta} - \beta_0) \\ n^{1/2}(\hat{\sigma}^2 - \sigma_0^2) \end{bmatrix} \underset{a}{\sim} N \left( \mathbf{0}, \begin{bmatrix} \sigma_0^2 \text{plim}(n^{-1}\mathbf{X}_0^\top \mathbf{X}_0)^{-1} & \mathbf{0} \\ \mathbf{0}^\top & 2\sigma_0^4 \end{bmatrix} \right). \quad (8.90)$$

Ce résultat et (8.89) sont tous deux corrects. Cependant, avec n'importe quel échantillon fini, l'intervalle de confiance pour  $\sigma$  basé sur (8.89) sera différent de l'intervalle de confiance basé sur (8.90). Comme nous en discuterons dans le Chapitre 13, le premier intervalle de confiance sera généralement plus précis, parce que la distribution de  $n^{1/2}(\hat{\sigma} - \sigma_0)$  sera plus proche de la distribution normale avec des échantillons finis que celle de  $n^{1/2}(\hat{\sigma}^2 - \sigma_0^2)$ . Il est alors préférable de paramétriser le modèle en termes de  $\sigma$  plutôt que de  $\sigma^2$ .

Dans la pratique, naturellement, nous sommes intéressés par  $\hat{\beta}$  et  $\hat{\sigma}$  plutôt que par  $n^{1/2}(\hat{\beta} - \beta_0)$  et  $n^{1/2}(\hat{\sigma} - \sigma_0)$ . Ainsi, au lieu d'utiliser (8.88), nous devrions en fait réaliser des inférences basées sur la matrice de covariance estimée

$$\hat{V}(\hat{\beta}, \hat{\sigma}) = \begin{bmatrix} \hat{\sigma}^2(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} & \mathbf{0} \\ \mathbf{0}^\top & \hat{\sigma}^2/2n \end{bmatrix},$$

dont le bloc supérieur gauche de dimension  $k \times k$  est l'estimateur NLS habituel de la matrice de covariance pour  $\hat{\beta}$ .

Dans la Section 8.1, nous avons considéré un exemple simple, (8.01), qui ne pouvait pas être estimé par moindres carrés. Si nous formulons l'hypothèse additionnelle que les aléas sont normalement distribués, ce modèle devient

$$y_t^\gamma = \beta_0 + \beta_1 x_t + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad (8.91)$$

qui ressemble presque à un modèle de régression, excepté que la variable dépendante est soumise à une transformation non linéaire.

La fonction de logvraisemblance correspondant à (8.91) est

$$\begin{aligned} \ell(\beta, \gamma, \sigma) = & -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^n (y_t^\gamma - \beta_0 - \beta_1 x_t)^2 \\ & + n \log|\gamma| + (\gamma - 1) \sum_{t=1}^n \log(y_t). \end{aligned} \quad (8.92)$$

Les trois premiers termes constituent exactement la fonction de logvraisemblance que nous obtiendrions si nous traitions  $y_t^\gamma$  comme la variable dépendante. Les quatrième et cinquième termes ne représentent en fait qu'un seul

terme, un terme Jacobien. Ce terme apparaît parce que  $\partial u_t / \partial y_t = \gamma y_t^{\gamma-1}$ . Par conséquent la contribution à la fonction de vraisemblance apportée par observation  $t$  doit inclure le facteur Jacobien  $|\gamma y_t^{\gamma-1}|$ , qui est la valeur absolue de  $\partial u_t / \partial y_t$ . En sommant sur tous les  $t$  et opérant le logarithme nous obtenons le terme qui apparaît dans (8.92).

En concentrant la fonction de logvraisemblance par rapport à  $\sigma$  nous aboutissons à

$$\begin{aligned} \ell^c(\boldsymbol{\beta}, \gamma) = & C - n \log \left( \sum_{t=1}^n (y_t^\gamma - \beta_0 - \beta_1 x_t)^2 \right) \\ & + n \log |\gamma| + (\gamma - 1) \sum_{t=1}^n \log(y_t). \end{aligned} \quad (8.93)$$

La maximisation de cette quantité par rapport à  $\gamma$  et  $\boldsymbol{\beta}$  est simple. Si un programme d'optimisation non linéaire convenable n'est pas disponible, on peut simplement faire une recherche à une dimension sur  $\gamma$ , en calculant  $\beta_0$  et  $\beta_1$  conditionnels à  $\gamma$  à l'aide des moindres carrés, afin de trouver la valeur  $\hat{\gamma}$  qui maximise (8.93). Naturellement, on ne peut pas utiliser la matrice de covariance OLS obtenue de cette manière, car elle traite l'estimation  $\hat{\gamma}$  comme fixée. La matrice d'information *n'est pas* bloc-diagonale entre  $\boldsymbol{\beta}$  et les autres paramètres de (8.91), aussi doit-on calculer et inverser la matrice d'information entière pour obtenir une matrice de covariance estimée.

L'estimation ML s'applique dans ce cas à cause du terme Jacobien qui apparaît dans (8.92) et (8.93). Il disparaît quand  $\gamma = 1$  mais joue un rôle extrêmement important pour toutes les autres valeurs de  $\gamma$ . Nous avons vu dans la Section 8.1 que si l'on appliquait les NLS à (8.01) et si tous les  $y_t$  étaient supérieurs à l'unité, on aboutirait à une estimation de  $\gamma$  infiniment grande et négative. Cela n'arrivera pas si l'on utilise le maximum de vraisemblance, parce que le terme  $(\gamma-1) \sum_{t=1}^n \log(y_t)$  ne tendra pas vers moins l'infini quand  $\gamma \rightarrow \infty$  beaucoup plus vite que le logarithme du terme de la somme des carrés ne tend vers plus l'infini. Cet exemple illustre l'utilité de l'estimation ML pour traiter des modèles de régression modifiés dans lesquels la variable dépendante est soumise à une transformation. Nous rencontrerons d'autres problèmes de ce type dans le Chapitre 14.

L'estimation ML peut aussi être très utile lorsque l'on croit que les aléas sont non normaux. Comme exemple extrême, considérons le modèle suivant:

$$y_t = \mathbf{X}_t \boldsymbol{\beta} + \alpha \varepsilon_t, \quad f(\varepsilon_t) = \frac{1}{\pi(1 + \varepsilon_t^2)}, \quad (8.94)$$

où  $\boldsymbol{\beta}$  est un vecteur de dimension  $k$  et  $\mathbf{X}_t$  est la  $t^{\text{ième}}$  ligne d'une matrice de dimension  $n \times k$ . La densité de  $\varepsilon_t$  est ici la densité de Cauchy (consulter la Section 4.6) et  $\varepsilon_t$  n'a donc pas de moments finis. Le paramètre  $\alpha$  est

simplement un paramètre d'échelle, et *non pas* l'écart type des aléas; comme la distribution de Cauchy n'a pas de moments, les aléas *n'ont pas* d'écart type.

Si nous écrivons  $\varepsilon_t$  comme une fonction de  $y_t$ , nous trouvons que

$$\varepsilon_t = \frac{y_t - \mathbf{X}_t\boldsymbol{\beta}}{\alpha}.$$

Ainsi, la densité de  $y_t$  est

$$f(y_t) = \frac{1}{\pi\alpha} \left( 1 + \frac{(y_t - \mathbf{X}_t\boldsymbol{\beta})^2}{\alpha^2} \right)^{-1},$$

le facteur  $1/\alpha$  étant un facteur Jacobien. La contribution à la fonction de logvraisemblance de la  $t^{\text{ième}}$  observation est ainsi

$$-\log(\pi) - \log(\alpha) - \log\left(1 + \frac{(y_t - \mathbf{X}_t\boldsymbol{\beta})^2}{\alpha^2}\right),$$

et la fonction de logvraisemblance elle-même est

$$\ell(\boldsymbol{\beta}, \alpha) = -n \log(\pi) - n \log(\alpha) - \sum_{t=1}^n \log\left(1 + \frac{(y_t - \mathbf{X}_t\boldsymbol{\beta})^2}{\alpha^2}\right). \quad (8.95)$$

Les conditions du premier ordre pour  $\hat{\beta}_i$  peuvent être écrites comme

$$-2\hat{\alpha}^{-2} \sum_{t=1}^n \left( 1 + \frac{(y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}})^2}{\hat{\alpha}^2} \right)^{-1} (y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}}) X_{ti} = 0. \quad (8.96)$$

L'expression équivalente pour l'estimation ML avec des erreurs normales (c'est-à-dire OLS) est

$$-\hat{\sigma}^{-2} \sum_{t=1}^n (y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}}) X_{ti} = 0. \quad (8.97)$$

La différence entre les équations de vraisemblance (8.96) et (8.97) est frappante. La dernière indique qu'une somme *non pondérée* des résidus fois chacun des régresseurs doit être égale à zéro. La première indique qu'une somme *pondérée* des mêmes quantités doit être égale à zéro, avec des poids inversement reliés à la taille des résidus. La raison de ceci est que la distribution de Cauchy génère de nombreuses valeurs extrêmes. Il y aura en général de nombreux aléas très importants, et afin d'éviter qu'ils n'influencent trop les estimations, la procédure ML d'estimation de  $\hat{\boldsymbol{\beta}}$  leur attribue beaucoup moins de poids que ne le font les OLS. Ces estimations ML possèdent toutes les propriétés habituelles de convergence, de normalité asymptotique, et ainsi de suite. Par contraste, si l'on appliquait simplement les OLS au modèle

(8.94), les aléas extrêmement grands fréquemment générés par la distribution de Cauchy feraient en sorte que les estimations ne soient même pas convergentes. Le théorème de convergence habituel pour les moindres carrés ne s'applique pas ici parce que les  $\varepsilon_t$  n'ont pas de moments finis.

Parce que les équations de vraisemblance (8.96) dépendent des résidus, la valeur  $\hat{\alpha}$  affecte la valeur  $\hat{\beta}$  qui les résoud. Ainsi, il est nécessaire de les résoudre conjointement pour  $\hat{\beta}$  et  $\hat{\alpha}$ . Malheureusement, il existe en général de multiples solutions à ces équations; voir Reeds (1985). Ainsi, une grande quantité d'efforts doit être consacrée à localiser le maximum global de la fonction de logvraisemblance (8.95).

## 8.11 CONCLUSION

Ce chapitre a fourni une introduction à toutes les caractéristiques majeures de l'estimation par maximum de vraisemblance et des tests de spécification, que nous utiliserons à travers le reste de ce livre. Le Chapitre 9 de Cox et Hinkley (1974) fournit un traitement plus détaillé sur certains des sujets que nous avons couverts. Une autre référence utile est Rothenberg (1973). Dans les deux prochains chapitres, nous utiliserons certains résultats de ce chapitre, avec les résultats antérieurs des estimateurs NLS et IV, pour traiter des sujets variés qui préoccupent les économètres. Le Chapitre 9 traite de la méthode des moindres carrés généralisés que l'on considère à la fois comme un exemple d'estimation ML et comme une extension des moindres carrés. Le Chapitre 10 traite ensuite du sujet très important de corrélation en série. Le Chapitre 13 fournira un traitement beaucoup plus détaillé sur les trois statistiques de test classiques que ne le fit la Section 8.9 et introduira une régression artificielle, comparable à la régression de Gauss-Newton, que l'on pourra utiliser avec des modèles estimés par ML.

## TERMES ET CONCEPTS

borne de Cramér-Rao	fonction de logvraisemblance
calcul (d'un estimateur)	concentrée
contributions à la fonction de	fonction de vraisemblance
vraisemblance et à la fonction de	identification: asymptotique
logvraisemblance	et fortement asymptotique,
convergence des estimateurs de	asymptotique sur un espace
Type 1 et 2	paramétrique non compact, globale,
distribution asymptotique (d'un	locale
estimateur)	information dans l'observation $t$
distribution exponentielle	invariance (à la reparamétrisation)
efficacité asymptotique	matrice CG
égalité de la matrice d'information	matrice de covariance asymptotique
équations de vraisemblance	maximum de vraisemblance (ML)
espace paramétrique	matrice d'information: asymptotique,
estimateur convergent au taux $n^{1/2}$	empirique et moyenne espérée
estimateur de la matrice	matrice Hessienne (fonction de
d'information produit-extérieur-du-	logvraisemblance): moyenne
gradient (OPG)	empirique, asymptotique, et espérée
estimation et estimateur	normalité asymptotique
estimateur par maximum de	paramétrisation d'un modèle
vraisemblance de Type 1 et 2	propriétés: normalité asymptotique,
estimation par maximum de	efficacité, asymptotique, calcul,
vraisemblance (MLE): Type 1 et 2	convergence, invariance
estimateur par maximum de	reparamétrisation
vraisemblance, propriétés:	statistiques de test classiques
efficacité asymptotique, normalité	terme Jacobien
asymptotique, calcul, convergence,	test (LM) du multiplicateur de
invariance	Lagrange
estimateur quasi-ML (QML) ou	test de rapport de vraisemblance
pseudo-ML	test de Wald
fonction (vecteur score)	test score (forme score du test ML)
	vecteur gradient de la fonction de
	logvraisemblance (vecteur score)



# Chapitre 9

## Le Maximum de Vraisemblance et Les Moindres Carrés Généralisés

### 9.1 INTRODUCTION

Jusqu'à présent nous avons supposé que les erreurs relatives aux modèles de régression sont indépendamment distribuées avec une variance constante. C'est une hypothèse très forte, qui est souvent mise à mal dans la pratique. Dans ce chapitre, nous envisageons des techniques d'estimation qui permettent de la relâcher. Ce sont d'une part les **moindres carrés généralisés**, ou **GLS**, et les **moindres carrés généralisés non linéaires**, ou **GNLS**, et d'autre part des applications variées de la méthode du maximum de vraisemblance. Nous traitons les GLS et le ML ensemble parce que quand le ML est appliqué aux modèles de régression dont les erreurs sont normales, les estimateurs qui en résultent entretiennent des liens étroits avec les estimateurs GLS.

Le plan de ce chapitre est le suivant. Tout d'abord, dans la Section 9.2, nous relâchons l'hypothèse selon laquelle les aléas sont indépendamment distribués avec une variance constante. L'estimation ML des modèles de régression sans ces hypothèses se trouve être conceptuellement simple et très proche de la méthode des GNLS. Dans la Section 9.3, nous discutons de la géométrie des GLS, et considérons un cas particulier important dans lequel les estimations OLS et GLS sont identiques. Dans la Section 9.4, nous décrivons la manière d'utiliser une version de la régression de Gauss-Newton avec des modèles estimés par GNLS. Dans la Section 9.5, nous établissons un lien entre les GNLS et les **GNLS faisables**, et discutons d'un certain nombre de résultats fondamentaux concernant à la fois les GNLS et les GNLS faisables. La relation entre les GNLS et le ML est ensuite traitée dans la Section 9.6. Enfin, de la Section 9.7 à la Section 9.9, nous considérons les **modèles de régression non linéaire multivariée**. Bien que de tels modèles puissent souvent paraître très difficiles, et en premier lieu à cause de la notation complexe qui doit permettre de prendre en compte de nombreuses variables dépendantes entre elles, nous montrons qu'ils sont en fait assez simples à estimer à l'aide des GNLS ou du ML. Pour terminer, dans la Section 9.10, nous discutons des modèles qui traitent des données de panel et d'autres ensembles de données qui combinent

des observations chronologiques et des données en coupe transversale. Dans ce chapitre, nous ne discutons pas de ce qui est probablement l'application des GLS la plus communément rencontrée, à savoir l'estimation des modèles de régression avec corrélation en série. L'énorme littérature sur ce sujet sera le thème du Chapitre 10.

## 9.2 LES MOINDRES CARRÉS GÉNÉRALISÉS

Nous nous proposons de considérer dans cette section la classe des modèles

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Omega}), \quad (9.01)$$

où  $\boldsymbol{\Omega}$ , une matrice définie positive de dimension  $n \times n$ , est la matrice de covariance du vecteur des aléas  $\mathbf{u}$ . L'hypothèse de normalité peut naturellement être relâchée, mais nous la conservons pour le moment puisque nous voulons utiliser la méthode du maximum de vraisemblance. Dans certaines applications, la matrice  $\boldsymbol{\Omega}$  peut être connue. Dans d'autres, elle peut être connue seulement à une constante multiplicative près, ce qui permet d'écrire  $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{\Delta}$ , avec  $\boldsymbol{\Delta}$  une matrice connue de dimension  $n \times n$  et  $\sigma^2$  un scalaire positif inconnu. Dans la plupart des applications, seule la structure de  $\boldsymbol{\Omega}$  sera connue; nous pourrions par exemple savoir qu'elle provient d'un schéma particulier d'hétéroscédasticité ou de corrélation en série, et par conséquent qu'elle dépend dans un sens d'un certain nombre de paramètres. Nous nous intéresserons à ces trois cas.

La fonction de densité du vecteur  $\mathbf{u}$  est la fonction de densité normale multivariée

$$f(\mathbf{u}) = (2\pi)^{-n/2} |\boldsymbol{\Omega}|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{u}^\top \boldsymbol{\Omega}^{-1} \mathbf{u}\right). \quad (9.02)$$

Afin de passer de la fonction de densité du vecteur des aléas  $\mathbf{u}$  à celle du vecteur des variables dépendantes  $\mathbf{y}$ , nous remplaçons  $\mathbf{u}$  par  $\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})$  dans (9.02) et nous multiplions par la valeur absolue du déterminant de la matrice Jacobienne associée à la transformation qui exprime  $\mathbf{u}$  en termes de  $\mathbf{y}$ . L'usage de ce facteur Jacobien est l'analogue de ce que nous avons déjà réalisé dans la Section 8.10 avec les variables aléatoires scalaires. Pour les détails, consulter l'Annexe B. Dans ce cas, la matrice Jacobienne correspond à la matrice identité, et son déterminant est égal à un. En conséquence, la fonction de vraisemblance est

$$L^n(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Omega}) = (2\pi)^{-n/2} |\boldsymbol{\Omega}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}))^\top \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}))\right),$$

et la fonction log-vraisemblance est

$$\ell^n(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Omega}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Omega}| - \frac{1}{2} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}))^\top \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})). \quad (9.03)$$

Si la matrice  $\Omega$  est connue, il est clair que cette fonction peut être maximisée par la minimisation de la **somme généralisée des résidus au carré**

$$SSR(\beta | \Omega) = (\mathbf{y} - \mathbf{x}(\beta))^\top \Omega^{-1} (\mathbf{y} - \mathbf{x}(\beta)). \quad (9.04)$$

Ce problème de minimisation est celui résolu par les **moindres carrés non linéaires généralisés**, ou **GNLS**. En dérivant (9.04) par rapport à  $\beta$  et en annulant les dérivées, nous obtenons  $k$  conditions du premier ordre comparables à (2.04):

$$-2\mathbf{X}^\top(\tilde{\beta})\Omega^{-1}(\mathbf{y} - \mathbf{x}(\tilde{\beta})) = \mathbf{0}. \quad (9.05)$$

La résolution de ces équations donne  $\tilde{\beta}$ , qui est le vecteur à la fois des estimations ML et GNLS pour ce problème. Il est simple de prolonger la théorie asymptotique du Chapitre 5, pour montrer que

$$n^{1/2}(\tilde{\beta} - \beta_0) \overset{a}{\sim} N\left(\mathbf{0}, \text{plim}_{n \rightarrow \infty} (n^{-1}\mathbf{X}^\top(\beta_0)\Omega^{-1}\mathbf{X}(\beta_0))^{-1}\right), \quad (9.06)$$

où  $\beta_0$  est la valeur de  $\beta$  sous le DGP. Ce résultat implique que nous pouvons réaliser des inférences pour les estimations GNLS essentiellement de la même manière que nous les réalisons pour les estimations NLS.

Dans le cas linéaire où  $\mathbf{x}(\beta) = \mathbf{X}\beta$ , les conditions du premier ordre (9.05) deviennent

$$-2\mathbf{X}^\top\Omega^{-1}\mathbf{y} + 2\mathbf{X}^\top\Omega^{-1}\mathbf{X}\tilde{\beta} = \mathbf{0}.$$

Celles-ci peuvent être résolues analytiquement pour donner la formule standard de l'estimateur des **moindres carrés généralisés**, ou **GLS**,<sup>1</sup>

$$\tilde{\beta} = (\mathbf{X}^\top\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\Omega^{-1}\mathbf{y}. \quad (9.07)$$

Cependant, en pratique, on calcule rarement les estimations GLS en utilisant cette formule. Supposons que  $\eta$  soit une matrice de dimension  $n \times n$  telle que

$$\eta^\top\eta = \Omega^{-1}. \quad (9.08)$$

Il existe différentes manières d'obtenir une matrice  $\eta$  qui satisfasse (9.08) (voir l'Annexe A); on la choisit habituellement, mais pas nécessairement, triangulaire. Etant donnée  $\eta$ , il est possible de calculer les estimations GLS au moyen de la régression OLS

$$\eta\mathbf{y} = \eta\mathbf{X}\beta + \eta\mathbf{u}. \quad (9.09)$$

Cette régression possède des erreurs qui sont indépendantes et qui ont une variance constante unitaire, puisque

$$E(\eta\mathbf{u}\mathbf{u}^\top\eta^\top) = \eta\Omega\eta^\top = \eta(\eta^\top\eta)^{-1}\eta^\top = \eta\eta^{-1}(\eta^\top)^{-1}\eta^\top = \mathbf{I}_n,$$

<sup>1</sup> L'estimateur GLS est occasionnellement appelé **estimateur Aitken**, parce qu'il fut proposé par Aitken (1935).

où  $\mathbf{I}_n$  est la matrice identité d'ordre  $n$ . L'estimation OLS de  $\beta$  provenant de la régression (9.09) est

$$\tilde{\beta} = (\mathbf{X}^\top \boldsymbol{\eta}^\top \boldsymbol{\eta} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\eta}^\top \boldsymbol{\eta} \mathbf{y} = (\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{y},$$

qui est l'estimation GLS de (9.07).

Le cas dans lequel  $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{\Delta}$ , où  $\sigma^2$  est inconnue mais où  $\boldsymbol{\Delta}$  est connue est pratiquement le même cas que celui où  $\boldsymbol{\Omega}$  est connue. La fonction de log-vraisemblance (9.03) devient

$$\begin{aligned} \ell^n(\mathbf{y}, \beta, \boldsymbol{\Delta}, \sigma) = & -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2} \log |\boldsymbol{\Delta}| \\ & - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x}(\beta))^\top \boldsymbol{\Delta}^{-1} (\mathbf{y} - \mathbf{x}(\beta)). \end{aligned}$$

La concentration de cette fonction par rapport à  $\sigma^2$  produit la fonction de log-vraisemblance concentrée

$$\ell^c(\mathbf{y}, \beta, \boldsymbol{\Delta}) = C - \frac{1}{2} \log |\boldsymbol{\Delta}| - \frac{n}{2} \log \left( (\mathbf{y} - \mathbf{x}(\beta))^\top \boldsymbol{\Delta}^{-1} (\mathbf{y} - \mathbf{x}(\beta)) \right).$$

Evidemment, cette quantité peut être maximisée en minimisant la somme généralisée des résidus au carré

$$SSR(\beta | \boldsymbol{\Delta}) = (\mathbf{y} - \mathbf{x}(\beta))^\top \boldsymbol{\Delta}^{-1} (\mathbf{y} - \mathbf{x}(\beta)),$$

qui ressemble exactement à (9.04) sauf que  $\boldsymbol{\Delta}$  joue maintenant le rôle de  $\boldsymbol{\Omega}$ . Ainsi, lorsque l'on souhaite réaliser une estimation, le fait que  $\boldsymbol{\Omega}$  soit complètement connue ou qu'elle soit connue à une constante multiplicative près importe peu.

Nous avons vu que si la matrice de covariance  $\boldsymbol{\Omega}$  est connue, au moins à une constante multiplicative près, il est simple conceptuellement de trouver les estimations GLS ou GNLS. Cependant, procéder ainsi peut ne pas être si aisé dans la pratique si  $n$  est important et si  $\boldsymbol{\Omega}^{-1}$  ou  $\boldsymbol{\eta}$  doivent être calculées numériquement. Heureusement, lorsque  $\boldsymbol{\Omega}$  est connue, ou lorsque sa structure l'est, elle dépend habituellement d'un nombre relativement restreint de paramètres, et une fois que ceux-ci ont été spécifiés, il est souvent possible de trouver analytiquement  $\boldsymbol{\Omega}^{-1}$  et  $\boldsymbol{\eta}$ . Dans un nombre important de cas semblables, la forme de  $\boldsymbol{\eta}$  est telle qu'il est extrêmement aisé de prémultiplier  $\mathbf{y}$  et  $\mathbf{X}$  par cette matrice. Nous rencontrerons plusieurs exemples de ce type lorsque nous discuterons de la corrélation en série dans le Chapitre 10.

Considérons l'exemple simple suivant, dans lequel les aléas sont hétéroscédastiques mais non corrélés les uns des autres

$$E(u_t^2) = \sigma^2 w_t^\alpha, \quad E(u_t u_s) = 0 \text{ pour } t \neq s, \quad (9.10)$$

où  $w_t$  est une observation portant sur une variable exogène et  $\alpha$  est un paramètre. Ce type de spécification pourrait avoir du sens si  $w_t$  était une variable liée à l'échelle de la variable dépendante, telle que la taille de l'entreprise si la variable dépendante était le bénéfice. Dans ce cas la matrice  $\mathbf{\Omega}$  est diagonale, avec un  $t^{\text{ième}}$  élément diagonal égal à  $\sigma^2 w_t^\alpha$ . Ainsi, la matrice  $\mathbf{\Omega}^{-1}$  est également une matrice diagonale avec  $\sigma^{-2} w_t^{-\alpha}$  comme  $t^{\text{ième}}$  élément diagonal, et  $\boldsymbol{\eta}$  est une matrice diagonale avec  $\sigma^{-1} w_t^{-\alpha/2}$  comme  $t^{\text{ième}}$  élément diagonal. La fonction  $\sigma^2 w_t^\alpha$  est ce que l'on appelle parfois la **fonction scédastique**. De la même manière qu'une fonction de régression détermine l'espérance conditionnelle d'une variable aléatoire, une fonction scédastique détermine sa variance conditionnelle.

Dans ce cas, il est particulièrement facile de voir qu'il n'est pas nécessaire de connaître  $\sigma$  pour obtenir les estimations GLS, puisque le sous-espace engendré par les colonnes de  $\boldsymbol{\eta}\mathbf{X}$  ne change pas si nous multiplions  $\boldsymbol{\eta}$  par n'importe quelle constante. Pourvu que nous connaissions  $\alpha$ , nous pouvons exécuter la régression

$$\frac{y_t}{w_t^{\alpha/2}} = \sum_{i=1}^k \beta_i \frac{X_{ti}}{w_t^{\alpha/2}} + \text{résidu.} \quad (9.11)$$

Elle produira exactement les mêmes estimations GLS  $\tilde{\boldsymbol{\beta}}$  que la régression (9.09), qui est dans ce cas

$$\frac{y_t}{\sigma w_t^{\alpha/2}} = \sum_{i=1}^k \beta_i \frac{X_{ti}}{\sigma w_t^{\alpha/2}} + \text{résidu.}$$

De (9.11) nous pouvons facilement estimer  $\sigma$ ; l'estimation correspond simplement à l'estimation OLS de l'écart type de la régression. Ce type de procédure GLS, dans laquelle la régressande et les régresseurs sont simplement multipliés par des pondérations qui varient au travers des observations est souvent appelé **moindres carrés pondérés**. Ceci s'applique à chaque fois que les aléas sont hétéroscédastiques avec des variances connues à une constante multiplicative près et non corrélés les uns aux autres.

Evidemment, il n'existe pas de difficulté conceptuelle à l'estimation des modèles tels que (9.01) quand la matrice de covariance  $\mathbf{\Omega}$  est connue, et de même il n'existe pas de difficulté conceptuelle à prouver que ces estimations possèdent les mêmes propriétés que les estimations NLS dans un modèle correctement spécifié. Cependant, l'estimation de  $\boldsymbol{\beta}$  devient beaucoup plus difficile lorsque  $\mathbf{\Omega}$  n'est pas connue. Dans ce cas, deux manières de procéder existent: les GNLS faisables, procédure dans laquelle l'inconnue  $\mathbf{\Omega}$  est remplacée par une matrice qui l'estime de façon convergente, et le maximum de vraisemblance. Nous considérons ces techniques respectivement dans les Sections 9.5 et 9.6.

## 9.3 LA GÉOMÉTRIE DES GLS

Dans cette section, nous discutons brièvement de la géométrie des moindres carrés généralisés. Les valeurs ajustées de la régression GLS de  $\mathbf{y}$  sur  $\mathbf{X}$  sont

$$\mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{y}.$$

De là, la matrice qui projette  $\mathbf{y}$  sur  $\mathcal{S}(\mathbf{X})$  est dans ce cas

$$\mathbf{P}_X^\Omega \equiv \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1}. \quad (9.12)$$

La matrice de projection complémentaire est

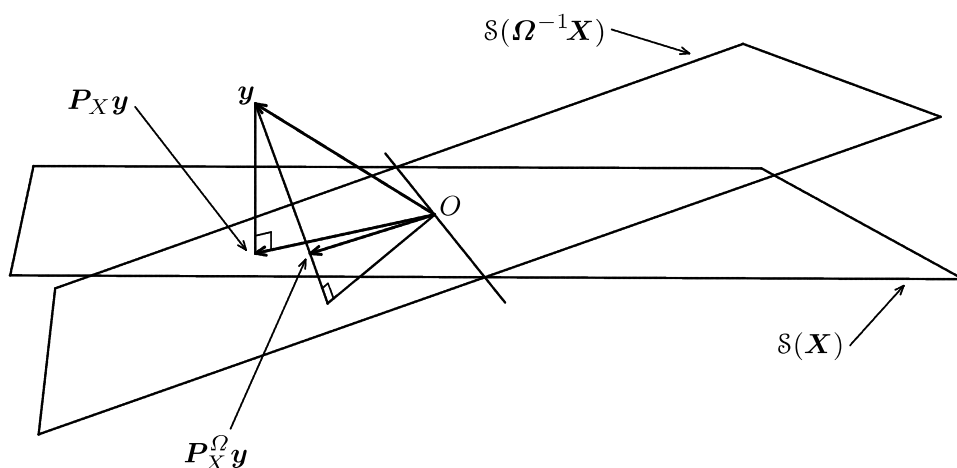
$$\mathbf{M}_X^\Omega \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1}. \quad (9.13)$$

Tout comme les matrices de projection les plus familières  $\mathbf{P}_X$  et  $\mathbf{M}_X$  associées aux moindres carrés ordinaires, il peut être facilement vérifié que ces matrices de projection sont idempotentes. Quoi qu'il en soit, comme elles ne sont pas symétriques,  $\mathbf{P}_X^\Omega$  ne projette pas *orthogonalement* sur  $\mathcal{S}(\mathbf{X})$ , et  $\mathbf{M}_X^\Omega$  projette sur  $\mathcal{S}^\perp(\boldsymbol{\Omega}^{-1} \mathbf{X})$  plutôt que sur  $\mathcal{S}^\perp(\mathbf{X})$ . Il existe des exemples où ces matrices sont appelées **matrices de projection oblique**, parce que l'angle entre les résidus  $\mathbf{M}_X^\Omega \mathbf{y}$  et les valeurs ajustées  $\mathbf{P}_X^\Omega \mathbf{y}$  n'est généralement pas égal à  $90^\circ$ . Pour s'en convaincre, observons que

$$\begin{aligned} \mathbf{y}^\top \mathbf{P}_X^\Omega \mathbf{M}_X^\Omega \mathbf{y} &= \mathbf{y}^\top \boldsymbol{\Omega}^{-1} \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1}) \mathbf{y} \\ &= \mathbf{y}^\top \boldsymbol{\Omega}^{-1} \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &\quad - \mathbf{y}^\top \boldsymbol{\Omega}^{-1} \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{y}, \end{aligned}$$

qui est nulle uniquement dans des circonstances très spéciales, telles que celles où  $\boldsymbol{\Omega}$  est proportionnelle à  $\mathbf{I}_n$ . Ainsi, les résidus des GLS ne sont généralement pas orthogonaux aux valeurs ajustées des GLS.

La Figure 9.1 illustre la distinction entre les estimations OLS et GLS. Dans le but d'avoir au plus trois dimensions dans nos représentations, quelques hypothèses simplificatrices ont dû être faites. Premièrement,  $\mathbf{X}$  et  $\boldsymbol{\Omega}^{-1} \mathbf{X}$  possèdent chacune seulement deux colonnes, afin que  $\mathcal{S}(\mathbf{X})$  et  $\mathcal{S}(\boldsymbol{\Omega}^{-1} \mathbf{X})$  puissent être bi-dimensionnelles. Ces deux sous-espaces sont représentés sur la figure par deux plans qui s'intersectent, mais en général, leur intersection se réduira seulement à l'origine. D'autre part, le vecteur  $\mathbf{y}$  appartient dans notre figure (par nécessité) au même espace à trois dimensions que les deux plans. En général, il n'en sera pas ainsi: normalement cinq dimensions sont nécessaires pour que la Figure 9.1 soit une représentation adéquate. Néanmoins, la figure est suffisante pour nos objectifs présents.



**Figure 9.1** Relation entre les estimations OLS et GLS

Les valeurs ajustées des OLS correspondent au vecteur  $P_X y$ , la projection orthogonale de  $y$  sur le plan  $S(X)$ . Afin de voir comment les résidus et les valeurs ajustées des GLS peuvent être construits géométriquement, souvenons-nous qu'à partir de (9.13) le champs de projection de  $M_X^\Omega$  est le complément orthogonal de  $S(\Omega^{-1}X)$ . Les résidus des GLS doivent alors se trouver dans  $S^\perp(\Omega^{-1}X)$ . D'un autre côté, les valeurs ajustées des GLS doivent se trouver dans  $S(X)$ , et ainsi  $y$  doit correspondre à la somme de deux vecteurs, non mutuellement orthogonaux, l'un appartenant à  $S(X)$  et l'un appartenant à  $S^\perp(\Omega^{-1}X)$ . Cette décomposition de  $y$  est illustrée sur la figure, sur laquelle nous pouvons voir directement que les résidus des GLS sont en réalité perpendiculaires à  $S(\Omega^{-1}X)$ .

Un autre point qui devrait être clair à partir de la figure est que le vecteur de résidus des GLS, en tant que résultat d'une projection oblique, doit nécessairement être plus long que le vecteur de résidus des OLS, qui est construit de manière à être le plus court possible. D'un autre côté, le vecteur des valeurs ajustées des GLS  $P_X^\Omega y$  peut être soit plus long soit plus court que le vecteur  $P_X y$  des valeurs ajustées des OLS. En fait, contrairement à  $P_X y$  qui est toujours plus court que  $y$ ,  $P_X^\Omega y$  peut être plus long que  $y$  dans certaines circonstances. La réalisation de l'une de ces possibilités dépend de la matrice de covariance  $\Omega$ . Pour un ensemble d'observations donné, il existe de nombreux ensembles différents d'estimations des GLS, un pour chaque choix possible de  $\Omega$ .

Nous pourrions en dire beaucoup plus concernant la géométrie des GLS et les propriétés des matrices de projection oblique; une référence classique est Seber (1980). Quoi qu'il en soit, ainsi que nous l'avons vu auparavant, la méthode des GLS est toujours équivalente à celle des OLS sur une régression dans laquelle la régressande et les régresseurs ont été convenablement transformés. Ainsi, tout ce que nous avons déjà appris concernant les OLS est

directement applicable aux GLS, dès que le modèle originel a été transformé comme dans (9.09). En particulier, le Théorème de Gauss-Markov s'applique aux modèles estimés par GLS. Si les données sont générées par un cas spécial de

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \boldsymbol{\Omega},$$

(notons que l'hypothèse de normalité n'est pas nécessaire ici), alors l'estimateur GLS (9.07) est le meilleur estimateur linéaire sans biais. Ce résultat découle de l'application du Théorème de Gauss-Markov démontré dans la Section 5.5 à la régression (9.09). De façon similaire, si le DGP est un cas particulier de (9.01) (peut-être avec  $\boldsymbol{\Omega} = \sigma^2\boldsymbol{\Delta}$  où seulement  $\boldsymbol{\Delta}$  est connue), alors l'estimateur GNLS sera le meilleur estimateur convergent et asymptotiquement linéaire.

Avant de quitter cette section, nous devons discuter de la possibilité importante où GLS et OLS peuvent dans certains cas donner des estimations identiques. Notre discussion fait suite à l'article de Kruskal (1968), et nous nous référerons alors au résultat en tant que **Théorème de Kruskal**. Le résultat est simple à énoncer: les estimations OLS et GLS sont les mêmes si et seulement si les deux sous-espaces  $\mathcal{S}(\mathbf{X})$  et  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  sont identiques. Le résultat est évident sur la Figure 9.1, imaginons simplement que  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  pivote pour coïncider avec  $\mathcal{S}(\mathbf{X})$ . Formellement, pour voir que les estimations OLS et GLS doivent coïncider si  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  et  $\mathcal{S}(\mathbf{X})$  sont les mêmes, il suffit d'observer que la décomposition par OLS de  $\mathbf{y}$  en un vecteur des valeurs ajustées et un vecteur de résidus satisfait les exigences de la décomposition (unique) par GLS:  $\mathbf{P}_X\mathbf{y}$  se trouve dans  $\mathcal{S}(\mathbf{X})$ , et  $\mathbf{M}_X\mathbf{y}$  est orthogonal à  $\mathcal{S}(\mathbf{X})$ , et par là aussi à  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$ . Si les valeurs ajustées par OLS  $\mathbf{X}\hat{\boldsymbol{\beta}}$  et les valeurs ajustées par GLS  $\mathbf{X}\tilde{\boldsymbol{\beta}}$  sont identiques, et si les estimations paramétriques  $\hat{\boldsymbol{\beta}}$  et  $\tilde{\boldsymbol{\beta}}$  sont uniques, ces deux procédures doivent être également identiques.

Le résultat réciproque, à savoir que si OLS et GLS donnent les mêmes estimations pour n'importe quelle réalisation du vecteur  $\mathbf{y}$ , alors  $\mathcal{S}(\mathbf{X})$  et  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  doivent être les mêmes, est aussi facile à voir. Notons qu'un unique vecteur de résidus doit être orthogonal à la fois à  $\mathcal{S}(\mathbf{X})$  et à  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$ , et par conséquent à  $\mathcal{S}(\mathbf{X}, \boldsymbol{\Omega}^{-1}\mathbf{X})$ . Puisque seuls  $k$  éléments de  $\boldsymbol{\beta}$  sont estimés, les résidus peuvent être orthogonaux à un espace à plus de  $k$  dimensions, et ainsi  $\mathcal{S}(\mathbf{X}, \boldsymbol{\Omega}^{-1}\mathbf{X})$ , peut être à plus de  $k$  dimensions. Mais comme  $\mathcal{S}(\mathbf{X})$  et  $\mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  sont tous deux de dimension  $k$ , ils doivent coïncider.

Selon les applications, il sera plus facile de manipuler  $\boldsymbol{\Omega}$  ou  $\boldsymbol{\Omega}^{-1}$ , et il peut être utile de noter que  $\mathcal{S}(\mathbf{X}) = \mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  si et seulement si  $\mathcal{S}(\mathbf{X}) = \mathcal{S}(\boldsymbol{\Omega}\mathbf{X})$ . Le raisonnement est comme suit:  $\mathcal{S}(\mathbf{X}) \subseteq \mathcal{S}(\boldsymbol{\Omega}^{-1}\mathbf{X})$  si et seulement si pour tout  $\boldsymbol{\beta} \in \mathbb{R}^k$  il existe  $\boldsymbol{\lambda} \in \mathbb{R}^k$  tel que  $\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\Omega}^{-1}\mathbf{X}\boldsymbol{\lambda}$ . Mais ceci est équivalent à dire que  $\boldsymbol{\Omega}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\lambda}$ , qui implique que  $\mathcal{S}(\boldsymbol{\Omega}\mathbf{X}) \subseteq \mathcal{S}(\mathbf{X})$ . La reprise de l'argumentation en permutant  $\mathbf{X}$  et  $\boldsymbol{\Omega}^{-1}\mathbf{X}$  donne le résultat dans son intégralité. La situation dans laquelle les estimations OLS et GLS sont identiques ne se rencontre pas très fréquemment, mais nous verrons une application importante du Théorème de Kruskal dans la Section 9.8.



Une autre manière de voir comment le Théorème de Kruskal peut être vérifié consiste à noter que l'estimateur GLS (9.07) peut être interprété comme un estimateur IV simple avec comme matrice d'instruments  $\Omega^{-1}\mathbf{X}$ . Nous savons de la Section 7.4 que l'estimateur IV simple est identique à l'estimateur IV généralisé. Ceci implique que

$$\tilde{\beta} = (\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Omega^{-1} \mathbf{y} = (\mathbf{X}^\top \mathbf{P}_{\Omega^{-1}\mathbf{X}} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_{\Omega^{-1}\mathbf{X}} \mathbf{y},$$

où, comme d'habitude,  $\mathbf{P}_{\Omega^{-1}\mathbf{X}}$  désigne la matrice de projection sur  $\mathcal{S}(\Omega^{-1}\mathbf{X})$ . Quand  $\mathcal{S}(\Omega^{-1}\mathbf{X}) = \mathcal{S}(\mathbf{X})$ ,  $\mathbf{P}_{\Omega^{-1}\mathbf{X}} = \mathbf{P}_\mathbf{X}$ . Ainsi, la seconde expression de  $\tilde{\beta}$  se réduit ici à l'expression de l'estimateur OLS  $\hat{\beta}$ .

Le fait que l'estimateur GLS ressemble à un estimateur IV suscite un intérêt plus théorique que pratique, parce que l'on ne voudrait pas obtenir des estimations GLS en utilisant une procédure IV. Les estimations paramétriques seraient correctes, mais l'estimation de la matrice de covariance ne le serait pas. La matrice de covariance correcte des GLS est proportionnelle à  $(\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1}$ , mais l'estimation IV de la matrice de covariance est proportionnelle à  $(\mathbf{X}^\top \mathbf{P}_{\Omega^{-1}\mathbf{X}} \mathbf{X})^{-1}$ .

#### 9.4 LA RÉGRESSION DE GAUSS-NEWTON

On associe à la méthode des GNLS une version de la régression de Gauss-Newton qui peut être utilisée dans des conditions identiques à l'utilisation de la régression de Gauss-Newton originelle (voir le Chapitre 6). Cette GNR est

$$\boldsymbol{\eta}(\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})) = \boldsymbol{\eta} \mathbf{X}(\boldsymbol{\beta}) \mathbf{b} + \text{résidus}, \quad (9.14)$$

où  $\mathbf{b}$  est un vecteur de coefficients à  $k$  dimensions qui doit être estimé et  $\boldsymbol{\eta}$  est n'importe quelle matrice de dimension  $n \times n$  qui satisfait l'équation (9.08). Ce n'est pas une coïncidence si la régression (9.14) ressemble à la régression (9.09), qui a été utilisée pour calculer les estimations GLS dans le cas linéaire. La GNR correspond en réalité à une linéarisation du modèle non linéaire originel, où à la fois la régressande et les régresseurs sont transformés afin de rendre la matrice de covariance des aléas proportionnelle à la matrice identité.

Si nous évaluons à la fois  $\mathbf{x}(\boldsymbol{\beta})$  et  $\mathbf{X}(\boldsymbol{\beta})$  en  $\tilde{\beta}$ , le résultat de la régression (9.14) donne  $\tilde{\mathbf{b}} = \mathbf{0}$  et la matrice de covariance estimée

$$\frac{(\mathbf{y} - \tilde{\mathbf{x}})^\top \boldsymbol{\eta}^\top \boldsymbol{\eta} (\mathbf{y} - \tilde{\mathbf{x}})}{n - k} (\tilde{\mathbf{X}}^\top \boldsymbol{\eta}^\top \boldsymbol{\eta} \tilde{\mathbf{X}})^{-1} = \frac{SSR(\tilde{\beta} | \Omega)}{n - k} (\tilde{\mathbf{X}}^\top \Omega^{-1} \tilde{\mathbf{X}})^{-1}. \quad (9.15)$$

Le premier facteur du membre de droite de (9.15) correspond précisément à l'estimation OLS de la variance de la GNR; comme nous l'expliquerons dans un moment, il doit tendre vers 1 quand  $n \rightarrow \infty$  si la matrice de covariance de  $\mathbf{u}$  est effectivement  $\Omega$ . Ce premier facteur serait normalement omis dans la

pratique.<sup>2</sup> En comparant le second facteur du membre de droite de (9.15) avec la matrice de covariance qui apparaît dans (9.06), il est évident que celui-ci fournit une estimation raisonnable de la matrice de covariance de  $\tilde{\beta}$ .

Au cours de la discussion précédente, nous faisons l'assertion selon laquelle  $(n - k)^{-1} SSR(\tilde{\beta} | \Omega)$  devrait tendre vers 1 lorsque  $n \rightarrow \infty$ . Avec cette assertion, nous avons utilisé implicitement le résultat suivant:

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \tilde{\mathbf{u}}^\top \Omega^{-1} \tilde{\mathbf{u}} \right) = 1. \quad (9.16)$$

Ce résultat demande justification. Tout d'abord, nous devons supposer que les valeurs propres de  $\Omega$ , qui sont toutes strictement positives puisque  $\Omega$  est supposée être définie positive, sont bornées supérieurement et inférieurement quand  $n \rightarrow \infty$ . Ces hypothèses impliquent que les valeurs de  $\eta$  possèdent les mêmes propriétés. Ensuite, nous utilisons le résultat selon lequel

$$\tilde{\mathbf{u}} = \mathbf{M}_0^\Omega \mathbf{u} + o(n^{-1/2}). \quad (9.17)$$

Ici,  $\mathbf{M}_0^\Omega$  désigne une matrice de projection oblique identique à (9.13), mais qui dépend de la matrice de dérivées  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$  plutôt que d'une matrice de régresseurs  $\mathbf{X}$ . Le résultat (9.17) est à l'évidence l'analogue GNLS du résultat (5.57) pour les NLS ordinaires et nous ne nous soucierons donc pas de le dériver.

Puisque l'hypothèse de la valeur bornée nous permet de conclure que

$$\eta \tilde{\mathbf{u}} = \eta \mathbf{M}_0^\Omega \mathbf{u} + o(n^{-1/2}),$$

la quantité dont nous voulons calculer la limite en probabilité dans (9.16) est

$$\begin{aligned} \frac{1}{n} \tilde{\mathbf{u}}^\top \Omega^{-1} \tilde{\mathbf{u}} &= \frac{1}{n} \left( \mathbf{u}^\top (\mathbf{M}_0^\Omega)^\top \Omega^{-1} \mathbf{M}_0^\Omega \mathbf{u} + o(n^{1/2}) \right) \\ &= \frac{1}{n} \mathbf{u}^\top (\mathbf{M}_0^\Omega)^\top \Omega^{-1} \mathbf{M}_0^\Omega \mathbf{u} + o(n^{-1/2}). \end{aligned} \quad (9.18)$$

Le premier terme dans la seconde ligne est ici

$$\begin{aligned} &\frac{1}{n} \mathbf{u}^\top (\mathbf{M}_0^\Omega)^\top \Omega^{-1} \mathbf{M}_0^\Omega \mathbf{u} \\ &= \frac{1}{n} \mathbf{u}^\top \Omega^{-1} \mathbf{u} - \frac{2}{n} \mathbf{u}^\top (\mathbf{P}_0^\Omega)^\top \Omega^{-1} \mathbf{u} + \frac{1}{n} \mathbf{u}^\top (\mathbf{P}_0^\Omega)^\top \Omega^{-1} \mathbf{P}_0^\Omega \mathbf{u} \\ &= \frac{1}{n} \mathbf{u}^\top \Omega^{-1} \mathbf{u} - \frac{1}{n} \mathbf{u}^\top \Omega^{-1} \mathbf{P}_0^\Omega \mathbf{u}, \end{aligned} \quad (9.19)$$

<sup>2</sup> Cet énoncé est vrai seulement si  $\Omega$  est complètement connue. Comme nous le verrons par la suite, l'estimateur GNLS demeure inchangé si  $\Omega$  est seulement connue à une constante multiplicative près, et il s'agit d'une estimation communément rencontrée dans la pratique. Dans ce cas, le premier facteur dans (9.15) serait employé pour estimer cette constante.

où

$$\mathbf{P}_0^\Omega \equiv \mathbf{I} - \mathbf{M}_0^\Omega \equiv \mathbf{X}_0(\mathbf{X}_0^\top \boldsymbol{\Omega}^{-1} \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \boldsymbol{\Omega}^{-1}$$

est essentiellement la même matrice que  $\mathbf{P}_X^\Omega$  définie dans (9.12). Seul le premier terme de (9.19) est  $O(1)$ . Intuitivement, la raison est que lorsque  $\mathbf{u}$  est projeté sur  $\mathcal{S}(\mathbf{X}_0)$ , le résultat se trouve dans un espace à  $k$  dimensions. Ainsi, une expression comparable au second terme dans (9.19), qui peut être écrite comme

$$n^{-1} (n^{-1/2} \mathbf{u}^\top \boldsymbol{\Omega}^{-1} \mathbf{X}_0) (n^{-1} \mathbf{X}_0^\top \boldsymbol{\Omega}^{-1} \mathbf{X}_0)^{-1} (n^{-1/2} \mathbf{X}_0^\top \boldsymbol{\Omega}^{-1} \mathbf{u}),$$

est  $O(n^{-1})$ , puisque chaque facteur sauf le premier est  $O(1)$ .

Ainsi, de (9.18) et (9.19), nous concluons que

$$\frac{1}{n} \tilde{\mathbf{u}}^\top \boldsymbol{\Omega}^{-1} \tilde{\mathbf{u}} \stackrel{a}{=} \frac{1}{n} \mathbf{u}^\top \boldsymbol{\Omega}^{-1} \mathbf{u}. \quad (9.20)$$

La forme quadratique dans le membre de droite de (9.20) peut être écrite très simplement en utilisant une matrice  $\boldsymbol{\eta}$  qui satisfait (9.08). Nous obtenons

$$\frac{1}{n} \mathbf{u}^\top \boldsymbol{\Omega}^{-1} \mathbf{u} = \frac{1}{n} \sum_{t=1}^n (\boldsymbol{\eta} \mathbf{u})_t^2.$$

Le vecteur  $\boldsymbol{\eta} \mathbf{u}$  possède une espérance nulle et une matrice de variance égale à  $\mathbf{I}_n$ . Les termes de la somme dans le membre de droite de cette expression sont alors non corrélés et asymptotiquement indépendants. Ainsi, nous pouvons appliquer une loi des grands nombres et affirmer que la limite en probabilité de la somme est égale à un. Il s'ensuit que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{u}^\top \boldsymbol{\Omega}^{-1} \mathbf{u} \right) = 1.$$

Alors, à partir de (9.20), nous concluons que cela reste vrai si  $\mathbf{u}$  est remplacé par  $\tilde{\mathbf{u}}$ , qui était ce que nous cherchions à montrer à l'origine.

Ce résultat peut être utilisé pour tester si  $\boldsymbol{\Omega}$  est réellement la matrice de covariance des aléas. Une statistique de test appropriée est  $\tilde{\mathbf{u}}^\top \boldsymbol{\Omega}^{-1} \tilde{\mathbf{u}}$ , qui correspond simplement à la SSR de la régression GNLS d'origine après transformation. Elle devrait être asymptotiquement distribuée suivant une  $\chi^2(n - k)$  sous l'hypothèse nulle.

## 9.5 LES MOINDRES CARRÉS GÉNÉRALISÉS FAISABLES

Dans la pratique, on connaît rarement la matrice de covariance  $\mathbf{\Omega}$ , mais on suppose parfois qu'elle dépend d'une manière particulière d'un vecteur  $\alpha$  de paramètres inconnus. Dans un tel cas, deux manières de procéder s'offrent à l'utilisateur. La première consiste à obtenir une estimation convergente de  $\alpha$ , disons  $\hat{\alpha}$ , par une quelconque procédure auxiliaire. Ceci produit alors une estimation de  $\mathbf{\Omega}$ ,  $\mathbf{\Omega}(\hat{\alpha})$ , qui est utilisée à la place de la véritable matrice de covariance  $\mathbf{\Omega}_0 \equiv \mathbf{\Omega}(\alpha_0)$  dans ce qui est en dehors de cette adaptation une procédure GLS standard. Cette approche, qui fera l'objet de cette section, est appelée **GLS faisables** parce qu'elle est faisable dans un grand nombre de cas où les GLS ordinaires ne le sont pas. L'autre approche consiste à utiliser le maximum de vraisemblance pour estimer  $\alpha$  et  $\beta$  conjointement, généralement sous l'hypothèse de normalité; cela sera discuté dans la Section 9.6.<sup>3</sup>

Sous des conditions raisonnables, les GLS faisables donnent des estimations qui non seulement sont convergentes mais aussi asymptotiquement équivalentes aux véritables estimations GLS, et par conséquent, elles partagent leurs propriétés d'efficacité. Cependant, même lorsque c'est le cas, les qualités des GLS faisables avec des échantillons finis peuvent être nettement amoindries par rapport à celle des véritables GLS si  $\hat{\alpha}$  est un estimateur pauvre de  $\alpha$ .

Dans la plupart des cas, les estimations de  $\alpha$  qui sont utilisées pour les GLS faisables sont basées sur les résidus OLS ou NLS, dont un élément type est  $\hat{u}_t \equiv y_t - x_t(\hat{\beta})$ . Il est envisageable d'utiliser ces résidus dans le but d'estimer  $\alpha$  parce que, dans de nombreuses circonstances, ils estiment de manière convergente les aléas  $u_t$ , bien qu'étant basés sur une procédure d'estimation qui utilise une matrice de covariance inappropriée. Il est évident que si les estimations OLS ou NLS  $\hat{\beta}$  estiment  $\beta$  de manière convergente, les résidus estimeront les aléas de manière convergente. Ce qui n'est pas évident (et n'est pas toujours vrai) est que  $\hat{\beta}$  estime de manière convergente  $\beta$ .

Un traitement rigoureux des conditions sous lesquelles les estimations NLS sont convergentes lorsque les aléas  $u_t$  ne satisfont pas l'hypothèse i.i.d. dépasse le domaine de ce livre. Consulter Gallant (1987) pour un tel traitement. Cependant, il est utile de voir comment la preuve de convergence de la Section 5.3 serait affectée si nous relâchions cette hypothèse. Souvenons-nous que la convergence de  $\hat{\beta}$  dépend entièrement des propriétés de  $n^{-1}$  fois la fonction somme-des-carrés:

$$ssr(\mathbf{y}, \beta) \equiv \frac{1}{n} \sum_{t=1}^n (y_t - x_t(\beta))^2 = \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta) + u_t)^2. \quad (9.21)$$

<sup>3</sup> Tout ceci suppose que la structure de  $\mathbf{\Omega}$  est connue. Lorsque ce n'est pas le cas, il n'est généralement pas possible d'utiliser les GNLS ou le ML. Cependant, comme nous le verrons dans le Chapitre 17, on peut tout de même obtenir des estimations qui sont plus efficaces que les estimations NLS en utilisant la méthode généralisée des moments.

Ici l'expression la plus à droite peut être réécrite comme

$$\frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta))^2 + \frac{2}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta))u_t + \frac{1}{n} \sum_{t=1}^n u_t^2. \quad (9.22)$$

Comme nous l'avons vu dans la Section 5.3, les trois termes de (9.22) doivent chacun satisfaire une propriété cruciale. Le premier terme doit satisfaire

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n (x_t(\beta_0) - x_t(\beta))^2 \right) > 0 \quad (9.23)$$

pour tout  $\beta \neq \beta_0$ . Cette propriété doit rester valable si le modèle doit être identifié asymptotiquement, et nous supposons qu'il l'est. Evidemment la condition (9.23) dépend seulement de la spécification de la fonction de régression, et non pas de l'éventuelle propriété i.i.d. des  $u_t$ , et il n'est donc pas nécessaire de nous y intéresser par la suite.

La seconde propriété cruciale est que le second terme de (9.22) doit tendre asymptotiquement vers zéro. Cette propriété dépend à l'évidence des propriétés des aléas  $u_t$ . S'ils sont indépendants, même s'ils ne sont pas identiquement distribués, alors l'argument de la Section 5.3 s'applique tel quel inchangé et montre que ce second terme a une espérance nulle. A condition que les variances des  $u_t$  et des fonctions de régression  $x_t(\beta)$  soient convenablement bornées, la loi des grands nombres pour les martingales, Théorème 4.6, peut être appliquée, et nous obtenons le résultat désiré. Pourtant, si les  $u_t$  ne sont pas indépendants, et si  $x_t(\beta)$  dépend des variables dépendantes retardées, il est très probable que le second terme de (9.22) n'aura pas une espérance nulle. Evidemment, nous devons écarter la combinaison dangereuse d'une fonction de régression qui dépend des variables dépendantes retardées et d'aléas qui sont dépendants entre eux. En règle générale, nous devons également écarter des aléas  $u_t$  dont les variances sont potentiellement infinies si nous désirons employer les lois des grands nombres.

La troisième propriété cruciale est que le dernier terme de (9.22) devrait avoir une limite en probabilité déterministe. Dans le cas i.i.d, il tend vers  $\sigma_0^2$ . Si les  $u_t$  sont indépendants mais non nécessairement identiquement distribués, cette propriété restera valable si la limite de la variance des erreurs *moyenne* existe. Une fois de plus nous devons en général écarter les variances potentiellement non bornées. Mais la propriété peut faire défaut si les  $u_t$  manifestent une trop forte corrélation entre eux. A titre d'exemple, supposons que les  $u_t$  soient identiquement distribués mais **équicorrélés**, ce qui signifie que la corrélation entre  $u_t$  et  $u_s$  est la même pour tout  $t \neq s$ . Ceci implique que nous pouvons écrire

$$u_t = \delta v + e_t, \quad (9.24)$$

pour un quelconque paramètre  $\delta$ , où  $v$  et  $e_t$  sont des variables aléatoires indépendantes, chacune de variance  $\omega^2$ . De là

$$E(u_t^2) = (\delta^2 + 1)\omega^2 \equiv \sigma^2$$

et, pour tout  $t \neq s$ ,

$$E(u_t u_s) = \delta^2 \omega^2.$$

Il s'ensuit que la corrélation entre  $u_t$  et  $u_s$  est  $\delta^2/(\delta^2 + 1)$ . En faisant varier  $\delta$ , nous pouvons évidemment donner à cette corrélation n'importe quelle valeur comprise entre zéro et un.

Le point clef de cet exemple réside dans la relation (9.24). En substituant celle-ci dans le troisième terme de (9.22), nous obtenons

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n u_t^2 &= \frac{1}{n} \sum_{t=1}^n (\delta v + e_t)^2 \\ &= \frac{1}{n} \sum_{t=1}^n (\delta^2 v^2 + 2\delta e_t v + e_t^2) \\ &= \delta^2 v^2 + \frac{1}{n} \sum_{t=1}^n (2\delta e_t v + e_t^2). \end{aligned}$$

Si nous travaillons conditionnellement à  $v$ , le second terme de la dernière expression précédente satisfait la loi des grands nombres la plus simple et tend vers la limite en probabilité déterministe égale à  $\omega^2$ . Mais le premier terme, qui est indépendant de la taille de l'échantillon, correspond à une variable aléatoire non dégénérée. Il en résulte que  $n^{-1}$  fois la fonction somme-des-carrés, l'expression (9.21), ne sera pas asymptotiquement non stochastique, et les estimations NLS  $\hat{\beta}$  ne seront pas convergentes.

Nous revenons maintenant au sujet des GLS faisables. Si nous pouvons éliminer la possibilité de variances non bornées, une dépendance en série beaucoup trop forte (similaire au cas pathologique que nous venons juste de décrire), et la combinaison de la corrélation en série et des variables dépendantes retardées dépendantes, les estimations NLS  $\hat{\beta}$  seront convergentes ainsi que les résidus  $\hat{u}_t$ . Nous pouvons alors utiliser ces résidus pour obtenir des estimations convergentes au taux  $n^{1/2}$  des paramètres  $\alpha$ . La méthode des GLS faisables s'applique à chaque fois que nous pouvons éliminer de telles éventualités.

A titre d'exemple considérons (9.10). Selon ce modèle, la variance de  $u_t$  est  $\sigma^2 w_t^\alpha$ , qui dépend des paramètres inconnus  $\alpha$  et  $\sigma^2$ . Une manière d'estimer  $\alpha$  consiste à exécuter la régression non linéaire

$$\hat{u}_t^2 = \sigma^2 w_t^\alpha + \text{residu}. \quad (9.25)$$

A condition que  $\hat{u}_t^2$  estime effectivement  $u_t^2$  de manière convergente, il semble hautement plausible que l'estimation NLS  $\hat{\alpha}$  à partir de (9.25) fournira une estimation convergente au taux  $n^{1/2}$  de  $\alpha$ . C'est un cas inhabituellement délicat puisque la régression auxiliaire qui permet d'estimer le paramètre de la matrice de covariance,  $\alpha$ , est non linéaire. Une autre manière d'estimer  $\alpha$

sera présentée dans la prochaine section. Nous rencontrerons certains cas plus simples, où les paramètres de la matrice de covariance peuvent être estimés par moindres carrés ordinaires, au cours du Chapitre 10.

Nous présentons maintenant une explication non rigoureuse de l'équivalence asymptotique entre les GNLS faisables et les GNLS. Les conditions du premier ordre pour GNLS sont

$$-2\mathbf{X}^\top(\tilde{\beta})\boldsymbol{\Omega}_0^{-1}(\mathbf{y} - \mathbf{x}(\tilde{\beta})) = \mathbf{0}. \quad (9.26)$$

Les conditions du premier ordre pour les GNLS faisables sont

$$-2\mathbf{X}^\top(\check{\beta})\check{\boldsymbol{\Omega}}^{-1}(\mathbf{y} - \mathbf{x}(\check{\beta})) = \mathbf{0}, \quad (9.27)$$

où  $\check{\beta}$  désigne l'estimateur des GNLS faisables et  $\check{\boldsymbol{\Omega}} \equiv \boldsymbol{\Omega}(\check{\alpha})$ . Evidemment, ces deux ensembles de conditions du premier ordre semblent en effet très similaires; la seule différence étant que  $\boldsymbol{\Omega}^{-1}$  apparaît dans (9.26) et  $\check{\boldsymbol{\Omega}}^{-1}$  apparaît dans (9.27). Mais comme  $\check{\alpha}$  est supposé être convergent au taux  $n^{1/2}$ , et  $\boldsymbol{\Omega}$  est supposée dépendre de et être dérivable par rapport à  $\alpha$ , nous pouvons écrire

$$\check{\boldsymbol{\Omega}}^{-1} = \boldsymbol{\Omega}_0^{-1} + \mathbf{A}, \quad \mathbf{A} = O(n^{-1/2}). \quad (9.28)$$

Par cette notation, nous signifions que chaque élément de la matrice  $\mathbf{A}$  est  $O(n^{-1/2})$ , ce qui implique que chaque élément de  $\check{\boldsymbol{\Omega}}^{-1}$  diffère de l'élément correspondant de  $\boldsymbol{\Omega}_0^{-1}$  d'une quantité qui est asymptotiquement négligeable. De là, (9.27) devient

$$-2\mathbf{X}^\top(\check{\beta})\boldsymbol{\Omega}_0^{-1}(\mathbf{y} - \mathbf{x}(\check{\beta})) - 2\mathbf{X}^\top(\check{\beta})\mathbf{A}(\mathbf{y} - \mathbf{x}(\check{\beta})) = \mathbf{0}. \quad (9.29)$$

Comme  $\boldsymbol{\Omega}_0$  est  $O(1)$  alors que  $\mathbf{A}$  est  $O(n^{-1/2})$ , le second terme ici devient négligeable relativement au premier lorsque  $n \rightarrow \infty$ . Mais le premier terme est simplement le membre de gauche de (9.26). Ainsi, asymptotiquement, les équations qui définissent l'estimateur des GNLS faisables  $\check{\beta}$  sont les mêmes que celles qui définissent l'estimateur GNLS  $\tilde{\beta}$ . Par conséquent, les deux estimateurs sont asymptotiquement équivalents.

Nous insistons sur le fait que la discussion précédente n'est pas rigoureuse. Nous n'avons pas montré formellement qu'il est correct d'écrire (9.28), ou que le second terme du membre de gauche de (9.29) est asymptotiquement négligeable relativement au premier. Cependant, une preuve pleinement rigoureuse de l'équivalence asymptotique des estimations des GLS et des GLS faisables est assez technique et pas très intuitive. Consulter Amemiya (1973a, 1973b) et Carroll et Ruppert (1982), parmi d'autres.

En pratique, le désir d'utiliser les GLS faisables comme méthode d'estimation dépend de la qualité de l'estimation de  $\boldsymbol{\Omega}$  que l'on peut obtenir. Si  $\boldsymbol{\Omega}(\check{\alpha})$  est une très bonne estimation de  $\boldsymbol{\Omega}_0$ , alors les GLS faisables auront, en effet,

essentiellement les mêmes propriétés que les GLS, et les inférences basées sur la matrice de covariance habituelle

$$(\check{\mathbf{X}}^\top \check{\boldsymbol{\Omega}}^{-1} \check{\mathbf{X}})^{-1} \quad (9.30)$$

seront raisonnablement fiables. Quoi qu'il en soit, si  $\boldsymbol{\Omega}(\check{\boldsymbol{\alpha}})$  est une estimation pauvre de  $\boldsymbol{\Omega}_0$ , les estimations des GLS faisables peuvent posséder des propriétés très différentes des véritables estimations GLS, et (9.30) peut mener à des inférences très trompeuses.

## 9.6 LE MAXIMUM DE VRAISEMBLANCE ET LES GNLS

Une seconde approche, qui est largement utilisée à la place des GLS faisables lorsque l'on suppose que  $\boldsymbol{\Omega}$  est donnée par  $\boldsymbol{\Omega}(\boldsymbol{\alpha})$  où  $\boldsymbol{\alpha}$  est inconnu, consiste à utiliser la méthode du maximum de vraisemblance. Pour l'utiliser, nous devons formuler une hypothèse concernant la distribution des aléas (dans la pratique, presque toujours de normalité). Ceci nous permet de noter la fonction de logvraisemblance appropriée comme une fonction du vecteur  $\boldsymbol{\alpha}$  de dimension  $q$  et du vecteur  $\boldsymbol{\beta}$  de dimension  $k$ .

Considérons la classe des modèles

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\alpha})). \quad (9.31)$$

En modifiant légèrement la fonction de logvraisemblance (9.03), nous trouvons que la fonction de logvraisemblance correspondant à (9.31) est

$$\begin{aligned} \ell^n(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = & -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Omega}(\boldsymbol{\alpha})| \\ & - \frac{1}{2} (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta}))^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha}) (\mathbf{y} - \mathbf{x}(\boldsymbol{\beta})). \end{aligned} \quad (9.32)$$

Deux ensembles de conditions du premier ordre existent, un pour  $\boldsymbol{\alpha}$  et un pour  $\boldsymbol{\beta}$ . Le second sera similaire aux conditions du premier ordre (9.05) pour les GNLS:

$$-2\mathbf{X}^\top(\hat{\boldsymbol{\beta}}) \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\alpha}}) (\mathbf{y} - \mathbf{x}(\hat{\boldsymbol{\beta}})) = \mathbf{0}.$$

Le premier sera plutôt compliqué, et dépendra précisément des liens entre  $\boldsymbol{\Omega}$  et  $\boldsymbol{\alpha}$ . Pour un traitement plus détaillé, consulter Magnus (1978).

Dans la Section 8.10, nous avons vu que la matrice d'information pour  $\boldsymbol{\beta}$  et  $\sigma$  dans un modèle de régression non linéaire avec pour matrice de covariance  $\sigma^2 \mathbf{I}$  est bloc-diagonale entre  $\boldsymbol{\beta}$  et  $\sigma$ . Un résultat analogue se révèle être exact pour le modèle (9.31) également: la matrice d'information est bloc-diagonale entre  $\boldsymbol{\beta}$  et  $\boldsymbol{\alpha}$ . Ceci signifie que, asymptotiquement, les vecteurs  $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$  et  $n^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$  sont indépendants. Ainsi, le fait que  $\hat{\boldsymbol{\alpha}}$  soit estimé conjointement avec  $\hat{\boldsymbol{\beta}}$  peut être ignoré, et  $\hat{\boldsymbol{\beta}}$  aura les mêmes propriétés



asymptotiques que l'estimateur des GNLS  $\tilde{\beta}$  et que l'estimateur des GNLS faisables  $\hat{\beta}$ .

L'argument précédent ne nécessite pas que les aléas  $u_t$  soient réellement normalement distribués. Tout ce dont nous avons besoin est que les vecteurs  $n^{1/2}(\hat{\beta} - \beta_0)$  et  $n^{1/2}(\hat{\alpha} - \alpha_0)$  soient asymptotiquement indépendants et  $O(1)$  sous n'importe quel DGP qui ait vraiment généré les données. On montre que ceci est en fait le cas sous des conditions absolument générales, similaires aux conditions détaillées dans le Chapitre 5 pour que les moindres carrés soient convergents et asymptotiquement normaux; voir White (1982) et Gouriéroux, Monfort, et Trognon (1984) pour les résultats fondamentaux dans ce domaine. Comme nous l'avons vu dans la Section 8.1, lorsque la méthode du maximum de vraisemblance est appliquée à un ensemble de données pour lequel le DGP n'était pas en réalité un cas particulier du modèle estimé, l'estimateur qui en résulte est appelé un estimateur quasi-ML, ou estimateur QML. Naturellement dans la pratique presque tous les estimateurs ML que nous utilisons sont en réalité des estimateurs QML, puisque certaines hypothèses de nos modèles sont presque toujours fausses. Il est alors réconfortant de savoir que dans certaines situations fréquentes, dont celle-ci, les propriétés des estimateurs QML sont très similaires à celles des véritables estimateurs, malgré bien évidemment la perte de l'efficacité asymptotique.

Comme exemple concret d'estimation des GLS, des GLS faisables et du ML, considérons le modèle

$$\mathbf{y} = \mathbf{x}(\beta) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{\Omega}), \quad \Omega_{tt} = \sigma^2 w_t^\alpha, \quad \Omega_{ts} = 0 \quad \text{pour tout } t \neq s. \quad (9.33)$$

Ce modèle manifeste une hétéroscédasticité de la forme (9.10). Puisque le déterminant de  $\mathbf{\Omega}$  est

$$\sigma^{2n} \prod_{t=1}^n w_t^\alpha,$$

nous voyons à partir de (9.32) que la fonction de logvraisemblance est

$$\begin{aligned} \ell^n(\mathbf{y}, \beta, \alpha, \sigma) = & -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{\alpha}{2} \sum_{t=1}^n \log(w_t) \\ & - \sum_{t=1}^n \frac{(y_t - x_t(\beta))^2}{2\sigma^2 w_t^\alpha}. \end{aligned} \quad (9.34)$$

Si  $\alpha$  était connu, nous pourrions obtenir des estimations GNLS en estimant la régression non linéaire pondérée

$$\frac{y_t}{w_t^{\alpha/2}} = \frac{x_t(\beta)}{w_t^{\alpha/2}} + \frac{u_t}{w_t^{\alpha/2}}, \quad (9.35)$$

que nous connaissions  $\sigma$  ou pas. Les estimations NLS pondérées de (9.35) correspondraient aux estimations GNLS  $\tilde{\beta}$ . La régression de Gauss-Newton associée à (9.35) serait

$$\frac{1}{w_t^{\alpha/2}}(y_t - x_t(\beta)) = \frac{1}{w_t^{\alpha/2}}\mathbf{X}_t(\beta)\mathbf{b} + \text{résidu},$$

qui est un cas particulier de (9.14).

Si  $\alpha$  n'était pas connu, nous devrions utiliser soit les GNLS faisables, soit le ML. La difficulté avec la première méthode consiste à obtenir une estimation convergente de  $\alpha$  sans beaucoup trop d'effort. La première étape consiste à exécuter une régression non linéaire de  $\mathbf{y}$  sur  $\mathbf{x}(\beta)$ , en ignorant l'hétéroscédasticité des aléas, afin d'obtenir un ensemble de résidus par moindres carrés  $\tilde{\mathbf{u}}$  (nous utilisons la notation  $\tilde{\mathbf{u}}$  plutôt que la notation plus naturelle  $\hat{\mathbf{u}}$  parce que dans cette section, ce dernier désigne une estimation ML). Nous pouvons ensuite utiliser ces résidus pour estimer  $\alpha$ . Dans la section précédente, nous suggérions d'utiliser les moindres carrés non linéaires avec l'équation (9.25) pour mener à bien cette deuxième étape. Cette approche n'est pas nécessairement la meilleure. Le modèle (9.33) implique que

$$u_t^2 = \sigma^2 w_t^\alpha \varepsilon_t^2, \quad (9.36)$$

où  $\varepsilon_t$  est  $N(0, 1)$ . Cette spécification de la fonction scédastique n'incite pas en elle-même à utiliser les moindres carrés. En fait, le moyen le plus attrayant d'estimer  $\alpha$  consiste à prétendre que  $\tilde{u}_t$  est effectivement  $u_t$  à estimer  $\alpha$  à partir de (9.36) par le maximum de vraisemblance. Si nous remplaçons  $y_t - x_t(\beta)$  dans (9.34) par  $\tilde{u}_t$ , nous obtenons

$$\ell^n(\mathbf{y}, \alpha, \sigma) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{\alpha}{2} \sum_{t=1}^n \log(w_t) - \sum_{t=1}^n \frac{\tilde{u}_t^2}{2\sigma^2 w_t^\alpha}. \quad (9.37)$$

Il s'agit de la fonction de logvraisemblance pour  $\alpha$  et  $\sigma$  conditionnelle à  $\beta$ , correspondant au vecteur des estimations NLS  $\tilde{\beta}$ . La condition du premier ordre pour  $\sigma^2$  est

$$-\frac{n}{2\sigma^2} + \sum_{t=1}^n \frac{2w_t^\alpha \tilde{u}_t^2}{4\sigma^4 w_t^{2\alpha}} = 0,$$

et sa résolution donne

$$\check{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n \frac{\tilde{u}_t^2}{w_t^\alpha}.$$

La substitution de  $\check{\sigma}^2$  dans (9.37) produit alors la fonction de logvraisemblance concentrée

$$\ell^c(\mathbf{y}, \alpha) = C - \frac{n}{2} \log\left(\frac{1}{n} \sum_{t=1}^n \frac{\tilde{u}_t^2}{w_t^\alpha}\right) - \frac{\alpha}{2} \sum_{t=1}^n \log(w_t). \quad (9.38)$$

Celle-ci peut être maximisée par une recherche de  $\alpha$  en une dimension.

Notons que lorsque  $\alpha$  devient plus grand, les second et troisième termes dans  $\ell^c(\mathbf{y}, \alpha)$  varieront dans des directions opposées. Le second terme est une expression somme-des-carrés, tandis que le troisième est un terme Jacobien. Pour être plus concret, supposons que  $w_t > 1$  pour tout  $t$ . Lorsque  $\alpha$  devient plus grand, le second terme augmente (puisque chaque  $\tilde{u}_t^2$  sera divisé par un nombre plus important, et que somme des résidus au carré pondérée diminuera sans cesse), mais le troisième terme deviendra plus petit (car  $\sum \log(w_t)$ , qui sera positive, sera multipliée par un plus grand nombre négatif). De plus, nous pouvons montrer que quand  $\alpha$  est suffisamment proche de zéro, l'augmentation dans le second terme doit être plus importante que la baisse dans le troisième, et que quand  $\alpha$  est suffisamment grand, le constat doit être vrai. Il doit donc exister une valeur positive, et finie  $\tilde{\alpha}$  qui maximise (9.38). Cette valeur serait utilisée dans la régression non linéaire (9.35) pour obtenir les estimations de GNLS faisables  $\tilde{\beta}$ .

Pour obtenir les estimations ML  $(\hat{\alpha}, \hat{\beta})$ , nous devons maximiser (9.34). En concentrant (9.34) par rapport à  $\sigma^2$ , nous obtenons la fonction de logvraisemblance concentrée

$$\ell^c(\mathbf{y}, \beta, \alpha) = C - \frac{n}{2} \log \left( \frac{1}{n} \sum_{t=1}^n \frac{(y_t - x_t(\beta))^2}{w_t^\alpha} \right) - \frac{\alpha}{2} \sum_{t=1}^n \log(w_t). \quad (9.39)$$

Celle-ci peut être maximisée par rapport à  $\alpha$  et  $\beta$  conjointement, en utilisant un algorithme général pour l'optimisation numérique.<sup>4</sup> Elle peut être également maximisée en utilisant la combinaison d'une recherche à une dimension sur  $\alpha$  et sur  $\beta$  conditionnellement à  $\alpha$ . La première approche est probablement la plus attrayante si  $\mathbf{x}(\beta)$  est non linéaire, même si la seconde peut l'être si  $\mathbf{x}(\beta) = \mathbf{X}\beta$ , car l'estimation de  $\beta$  conditionnelle à  $\alpha$  ne nécessitera qu'une simple régression OLS. Dans le second cas, nous pouvons effectivement concentrer par rapport à  $\beta$  et réduire (9.39) à une fonction de  $\alpha$  seulement.

Toute la discussion précédente a postulé l'absence de relation entre les paramètres  $\beta$  de la fonction de régression et les paramètres  $\alpha$  qui déterminent  $\Omega(\alpha)$ , et ceci est généralement une hypothèse raisonnable. Cependant, il est certainement possible d'établir des modèles où une telle relation existe. Un exemple est le modèle

$$y_t = \beta_0 + \beta_1(x_t^{\beta_2} z_t^{\beta_3}) + u_t, \quad u_t \sim N(0, \sigma^2 x_t^{\beta_2} z_t^{\beta_3}).$$

<sup>4</sup> La plupart des algorithmes d'optimisation numérique généraux procède essentiellement de la même manière que les algorithmes pour les moindres carrés non linéaires découverts dans la Section 6.8. La différence majeure est que la régression de Gauss-Newton ne peut pas être utilisée pour déterminer dans quelle direction chercher à chaque itération majeure. Pour maximiser les fonctions de logvraisemblance, d'autres régressions artificielles, que l'on détaillera dans les Chapitres 13, 14 et 15, peuvent être utilisées à la place, bien qu'il existe des algorithmes pratiques qui n'utilisent pas les régressions artificielles pour ce propos. Consulter Cramer (1986).

Ici les paramètres  $\beta_2$  et  $\beta_3$  apparaissent à la fois dans la fonction de régression et dans la fonction scédastique. Alors la matrice d'information n'est assurément pas bloc-diagonale entre les paramètres de la première fonction de régression et ceux de la seconde. Dans un cas comme celui-ci, le maximum de vraisemblance peut facilement être utilisé pour estimer efficacement tous les paramètres, tandis que les techniques comme les GNLS faisables, qui tentent d'estimer les paramètres de la fonction de régression conditionnellement à ceux de la fonction scédastique, n'en sont pas capables.

## 9.7 INTRODUCTION AUX RÉGRESSIONS MULTIVARIÉES

Jusqu'ici, et bien que nous ayons parfois donné formellement la possibilité à la variable dépendante dans les modèles que nous avons traités d'être un vecteur plutôt qu'un scalaire, nous n'avons effectivement pas discuté d'un quelconque modèle pour lequel c'est le cas. A présent nous sommes familiarisés avec les moindres carrés généralisés et avec l'utilisation du maximum de vraisemblance pour estimer les modèles de régression, et nous sommes prêts à discuter du **modèle de régression non linéaire multivariée**

$$y_{ti} = \xi_{ti}(\beta) + u_{ti}, \quad t = 1, \dots, n; \quad i = 1, \dots, m. \quad (9.40)$$

Ici  $y_{ti}$  est la  $t^{\text{ième}}$  observation de la  $i^{\text{ième}}$  variable dépendante,  $\xi_{ti}(\beta)$  est la  $t^{\text{ième}}$  observation de la fonction de régression qui détermine l'espérance conditionnelle de cette variable dépendante,  $\beta$  est un vecteur de dimension  $k$  regroupant les paramètres à estimer, et  $u_{ti}$  est un aléa d'espérance nulle et comportant d'autres propriétés dont nous discuterons dans peu de temps.

Les modèles de régression multivariée surviennent dans plusieurs circonstances. Comme exemple simple, supposons qu'il y ait des observations sur une variable dépendante, pour 5 pays sur 120 trimestres (ce qui implique que  $m = 5$  et  $n = 120$ ). Chaque pays pourrait avoir une fonction de régression différente déterminant l'espérance conditionnelle de la variable dépendante. Si les mêmes paramètres apparaissaient dans plus d'une fonction de régression, on dirait que le système est soumis à des **restrictions croisées**. En présence de telles restrictions, il est évident que l'on voudrait estimer les cinq équations simultanément dans un système plutôt qu'individuellement, afin d'obtenir des estimations efficaces. Même en l'absence de restrictions croisées, il semble très probable que les caractéristiques observées des environnements économiques des différents pays seraient reliées à chaque instant, de telle sorte que, selon toute vraisemblance,  $u_{ti}$  serait corrélé avec  $u_{tj}$  pour  $i \neq j$ . Dans cette situation, le système des équations forme un ensemble que Zellner (1962) surnomme **régressions sans lien apparent**, ou **système SUR**. En vérité, il semblerait plus logique de s'y référer en tant que "régressions avec lien apparent", mais il est trop tard pour changer la terminologie à ce stade. Comme Zellner l'a montré, l'estimation d'un ensemble de régressions sans lien apparent conjointement

dans un système produira sauf dans certains cas particuliers dont nous discuterons par la suite, des estimations plus efficaces que celles obtenues par l'estimation de chacune d'entre elles séparément, même quand il n'y a pas de restrictions croisées. Ainsi, nous voudrions normalement traiter un système SUR comme un modèle multivarié.

Il existe de nombreuses situations dans lesquelles la théorie économique suggère l'utilisation d'un modèle de régression multivariée. Une classe très largement répandue de modèles est celle des **systèmes de demande**, dans lesquels les parts des différentes classes de biens et services dans les dépenses des consommateurs sont reliées à la dépense totale et aux prix relatifs. La littérature sur les systèmes de demande est vaste; consulter, parmi de nombreux autres, Barten (1964, 1969, 1977), Brown et Heien (1972), Christensen, Jorgenson, et Lau (1975), Deaton (1974, 1978), Deaton et Muellbauer (1980), Parks (1969), Pollak et Wales (1969, 1978, 1981, 1987), Prais et Houthakker (1955), et Stone (1954). Les systèmes de demande peuvent être estimés en utilisant soit des données chronologiques agrégées (généralement annuelles mais parfois trimestrielles), ou, moins fréquemment, des données en coupe transversale ou un mélange de données chronologiques et de données en coupe transversale sur les ménages.

Dans bien des cas (bien que cela soit moins vrai dans la littérature plus récente), les formes fonctionnelles des systèmes de demande sont simplement obtenues en maximisant une fonction d'utilité d'une certaine forme connue soumise à une contrainte budgétaire. Par exemple, supposons que la fonction d'utilité soit

$$\sum_{i=1}^{m+1} \alpha_i \log(q_i - \gamma_i), \quad (9.41)$$

où il y a  $m + 1$  marchandises,  $q_i$  étant la quantité de marchandise  $i$  consommée et  $\alpha_i$  et  $\gamma_i$  étant des paramètres. La justification des  $m + 1$  marchandises apparaîtra bientôt. Les  $\alpha_i$  sont soumis à la restriction de normalisation  $\sum_{i=1}^{m+1} \alpha_i = 1$ .

La fonction d'utilité (9.41) est connue sous le nom de **fonction d'utilité Stone-Geary**. Sa maximisation sous la contrainte budgétaire

$$\sum_{i=1}^{m+1} q_i p_i = E,$$

où  $p_i$  est le prix de la marchandise  $i$  et  $E$  est la dépense totale pour toutes les marchandises, donne le système de demande:

$$s_i(E, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{\gamma_i p_i}{E} + \alpha_i \left( \frac{E - \sum_{j=1}^{m+1} p_j \gamma_j}{E} \right),$$

où  $s_i(E, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$  désigne la part de la dépense que l'on prévoit d'affecter à la marchandise  $i$ , conditionnelle à la dépense totale  $E$ , au vecteur de prix  $\mathbf{p}$  et aux

vecteurs de paramètres  $\alpha$  et  $\gamma$ . Ce système de demande particulier est connu sous le nom de **système de dépense linéaire**; cela relève d'une longue histoire antidatée de Stone (1954). Notons que bien que  $\alpha_i$  apparaisse seulement dans la  $i^{\text{ième}}$  équation de part,  $\gamma_i$  apparaît dans toutes les  $m + 1$  équations de part, de telle sorte qu'il existe un grand nombre de restrictions croisées.

Par définition, les parts dépensées sur toutes les marchandises doivent conduire à un total de un. Ceci constitue une implication importante pour les aléas, que nous n'avons pas encore spécifiés. Supposons que nous formulions l'hypothèse que

$$s_{ti} = s_i(E_t, \mathbf{p}_t, \alpha, \gamma) + u_{ti},$$

où  $s_{ti}$  est la part observée de la dépense correspondant à la marchandise  $i$  pour l'observation  $t$ , et  $u_{ti}$  est un aléa. Alors

$$\sum_{i=1}^{m+1} s_{ti} = \sum_{i=1}^{m+1} s_i(E_t, \mathbf{p}_t, \alpha, \gamma) + \sum_{i=1}^{m+1} u_{ti}.$$

Sommons les deux côtés de cette équation sur  $i$ , nous trouvons que  $1 = 1 + \sum_{i=1}^{m+1} u_{ti}$ , ce qui implique que

$$\sum_{i=1}^{m+1} u_{ti} = 0. \quad (9.42)$$

Ainsi les aléas pour chaque observation doivent avoir une somme nulle sur toutes les parts de la dépense. Comme Barten (1968) le montra, ceci ne crée pas de problème pour l'estimation; nous devons simplement abandonner une équation de part et estimer le système pour les  $m$  parts restantes. De plus, si nous utilisons le maximum de vraisemblance, le choix de l'équation que nous ne prenons pas en compte importe peu: les estimations de  $\alpha$  et  $\gamma$  que nous obtenons seront identiques (souvenons-nous que les  $\alpha_i$  sont normalisés pour donner une somme égale à l'unité; c'est pourquoi nous pouvons procéder sans avoir besoin d'estimer l'une d'entre elles).

Bien que (9.42) ne produise pas de problème sérieux pour l'estimation, elle laisse apparaître de manière absolument claire que les aléas  $u_{ti}$  et  $u_{tj}$  doivent être en général corrélés entre eux. A proprement parler, nous ne devrions pas supposer que les  $u_{ti}$  soient normalement distribués, parce que  $0 \leq s_{ti} \leq 1$ , ce qui implique que les  $u_{ti}$  doivent être borné supérieurement et inférieurement; voir Wales et Woodland (1983). Cependant, à condition que l'échantillon ne contienne pas d'observations qui sont, relativement aux écarts types de  $u_{ti}$ , proches de 0 ou 1, il est probablement raisonnable, en première approximation, de supposer la normalité, et c'est précisément ce que la plupart des auteurs ont fait. Ainsi, si  $\mathbf{U}_t$  désigne un vecteur ligne dont l'élément type est  $u_{ti}$ , nous pourrions spécifier la distribution des  $\mathbf{U}_t$  par  $N(\mathbf{0}, \Sigma)$ , où  $\Sigma$  est une matrice de covariance singulière de dimension  $(m + 1) \times (m + 1)$ . Alors

$$\mathbf{U}_t^* \sim N(\mathbf{0}, \Sigma^*),$$

où  $U_t^*$  correspond à  $U_t$  moins une composante, disons la dernière, et  $\Sigma^*$  est alors une sous-matrice de dimension  $m \times m$  de  $\Sigma$ . Parce que  $\Sigma$  est une matrice singulière, les systèmes d'équations pour lesquels la somme des aléas sur toutes les équations est nulle, sont fréquemment nommés **systèmes d'équations singuliers**; consulter Berndt et Savin (1975). Il existe de nombreux exemples de systèmes d'équations singuliers en plus des systèmes de demande. Ces systèmes incluent des systèmes de parts de facteurs de production tels que ceux décrits par (Berndt et Christensen (1974) et Fuss (1977), ainsi que des systèmes d'équations d'emploi-ressources (Aigner (1973)).

Nous retournons à présent aux modèles multivariés en général. La plus grande difficulté rencontrée avec de tels modèles est la notation. Comme  $\xi_{ti}(\beta)$  comporte déjà deux indices, ses dérivées premières et secondes par rapport aux éléments de  $\beta$  doivent avoir respectivement trois et quatre indices. Ceci rend difficile le traitement des modèles multivariés. L'utilisation de la notation matricielle conventionnelle n'est pas vraiment conçue pour manipuler des quantités avec plus de deux indices. La manière d'aborder le problème est propre à chaque auteur. À une extrémité, s'inspirant de la pratique de la physique moderne, Davidson et MacKinnon (1983b) préconisent l'utilisation de la "convention de la sommation d'Einstein", une notation qui évite en grande partie l'utilisation des matrices en traitant toutes les quantités comme des expressions scalaires impliquant (typiquement) plusieurs sommations sur les indices. Cette approche comporte de nombreux avantages. Malheureusement, bien qu'elle ait été utilisée par des économètres de grande notoriété, et parmi eux Sargan (1980b) et Phillips (1982), son utilisation n'est pas largement répandue en économétrie, et cela pourrait probablement sembler étrange à la plupart des lecteurs de ce livre. À l'autre extrémité, certains auteurs se livrent à une utilisation massive des produits de Kronecker ( $\otimes$ ), d'opérateurs vectoriels, et ainsi de suite, afin d'utiliser exclusivement la notation matricielle; consulter Magnus et Neudecker (1988). Comme Malinvaud (1970a), nous essaierons de tenir un cap intermédiaire, qui nous l'espérons, sera à la fois facile à comprendre et relativement facile à manipuler.

Puisque nous sommes dans des préoccupations de notation, remarquons que le modèle (9.40) pourrait être récrit sous l'un des deux formes suivantes:

$$Y_t = \xi_t(\beta) + U_t, \quad (9.43)$$

où  $Y_t$ ,  $\xi_t(\beta)$ , et  $U_t$  sont des vecteurs de dimension  $1 \times m$  avec des éléments types respectifs égaux à  $y_{ti}$ ,  $\xi_{ti}(\beta)$ , et  $u_{ti}$ , ou

$$Y = \xi(\beta) + U, \quad (9.44)$$

où  $Y$ ,  $\xi(\beta)$ , et  $U$  sont des matrices de dimension  $n \times m$  avec  $Y_t$ ,  $\xi_t(\beta)$ , et  $U_t$  comme lignes types. L'approche basée sur les conventions de sommation débiterait avec (9.40), tandis que l'approche basée sur les produits de Kronecker débiterait de (9.44), en utilisant les opérateurs "vec" et "vech" pour empiler les colonnes de  $Y$ ,  $\xi(\beta)$ , et  $U$ . Notre approche commencera à partir de (9.43).

## 9.8 L'ESTIMATION GLS DES RÉGRESSION MULTIVARIÉES

Dans la pratique, les modèles de régression multivariée sont estimés habituellement soit par les GLS faisables, soit par le maximum de vraisemblance, sous l'hypothèse de la normalité. Sauf dans des circonstances très rares, il n'est pas raisonnable de supposer que  $u_{ti}$  est indépendant de  $u_{tj}$  pour  $i \neq j$ , ainsi que nous l'avons déjà vu dans le cas à la fois des régressions sans lien apparent et des systèmes de demande. Selon que nous nous proposons d'utiliser le ML ou les GNLS faisables, nous pouvons ou pas vouloir supposer que le vecteur des aléas  $\mathbf{U}_t$  est normalement distribué. Dans l'un ou l'autre cas, nous ferons l'hypothèse que

$$\mathbf{U}_t \sim \text{IID}(\mathbf{0}, \boldsymbol{\Sigma}),$$

où  $\boldsymbol{\Sigma}$  est une matrice de covariance de dimension  $m \times m$  (habituellement inconnue) parfois désignée sous le nom de **matrice de covariance contemporaine**. Ainsi, nous supposons que  $u_{ti}$  est corrélé avec  $u_{tj}$  mais pas avec  $u_{sj}$  pour  $s \neq t$ . Ceci est bien entendu une hypothèse forte, qui devrait être testée; par la suite, nous discuterons d'un test qui peut parfois être approprié. Sous ces hypothèses, la somme généralisée des résidus au carré pour le modèle (9.43) est

$$\sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top. \quad (9.45)$$

Supposons pour l'instant que  $\boldsymbol{\Sigma}$  soit connue. Alors  $\boldsymbol{\Sigma}$  peut être utilisée pour transformer le modèle multivarié (9.40) en un modèle univarié. Supposons que  $\boldsymbol{\psi}$  soit une matrice de dimension  $m \times m$  (habituellement triangulaire) telle que

$$\boldsymbol{\psi} \boldsymbol{\psi}^\top = \boldsymbol{\Sigma}^{-1}. \quad (9.46)$$

Si nous postmultiplions chaque terme dans (9.43) par  $\boldsymbol{\psi}$ , nous obtenons la régression

$$\mathbf{Y}_t \boldsymbol{\psi} = \boldsymbol{\xi}_t(\boldsymbol{\beta}) \boldsymbol{\psi} + \mathbf{U}_t \boldsymbol{\psi}. \quad (9.47)$$

Le vecteur d'erreur de dimension  $1 \times m$   $\mathbf{U}_t \boldsymbol{\psi}$  a une matrice de covariance égale à

$$E(\boldsymbol{\psi}^\top \mathbf{U}_t^\top \mathbf{U}_t \boldsymbol{\psi}) = \boldsymbol{\psi}^\top \boldsymbol{\Sigma} \boldsymbol{\psi} = \mathbf{I}_m. \quad (9.48)$$

Comme nous l'avons écrit, (9.47) comporte seulement une observation, et tous les termes sont des vecteurs de dimension  $1 \times m$ . Afin de d'exécuter cette régression, nous devons d'une manière ou d'une autre convertir ces vecteurs de dimension  $1 \times m$  en des vecteurs de dimension  $nm \times 1$  regroupant toutes les observations. Il existe plus d'une manière de réaliser ceci.

Une approche consiste simplement de transposer chaque vecteur de dimension  $1 \times m$  de (9.47) et d'empiler ensuite les vecteurs de dimension  $m$  ainsi créés. Cependant, ceci n'est pas la manière la plus simple de procéder. Une approche plus facile est premièrement de former les  $m$  ensembles de vecteurs de dimension  $n$ , comme suit. Pour la variable dépendante, le  $t^{\text{ième}}$  élément



du  $i^{\text{ième}}$  vecteur serait  $\mathbf{Y}_t \boldsymbol{\psi}_i$ , où  $\boldsymbol{\psi}_i$  est la  $i^{\text{ième}}$  colonne de  $\boldsymbol{\psi}$ , et pour les fonctions de régression l'élément correspondant serait  $\boldsymbol{\xi}_t(\boldsymbol{\beta}) \boldsymbol{\psi}_i$ . Puis, les vecteurs de dimension  $nm$  seraient obtenus en empilant les vecteurs de dimension  $n$ . La régression non linéaire *univariée* ainsi définie peut être exprimée en termes des matrices partitionnées comme

$$\begin{bmatrix} \mathbf{Y} \boldsymbol{\psi}_1 \\ \vdots \\ \mathbf{Y} \boldsymbol{\psi}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi}(\boldsymbol{\beta}) \boldsymbol{\psi}_1 \\ \vdots \\ \boldsymbol{\xi}(\boldsymbol{\beta}) \boldsymbol{\psi}_m \end{bmatrix} + \begin{bmatrix} \mathbf{U} \boldsymbol{\psi}_1 \\ \vdots \\ \mathbf{U} \boldsymbol{\psi}_m \end{bmatrix}. \quad (9.49)$$

Souvenons-nous de (9.44) que  $\mathbf{Y}$ ,  $\boldsymbol{\xi}(\boldsymbol{\beta})$ , et  $\mathbf{U}$  sont toutes des matrices de dimension  $n \times m$ . La régression univariée empilée aura une matrice de covariance  $\mathbf{I}_{mn}$ , provenant de (9.48) et parce que nous avons supposé qu'il n'y a pas de corrélation non contemporaine des aléas. Même si  $\boldsymbol{\psi}$  était connue seulement à une constante multiplicative près, on pourrait estimer cette régression univariée par les moindres carrés non linéaires, tout comme n'importe quelle régression non linéaire univariée. En utilisant la notation de (9.47), sa somme des résidus au carré serait

$$\begin{aligned} & \sum_{t=1}^n (\mathbf{Y}_t \boldsymbol{\psi} - \boldsymbol{\xi}_t(\boldsymbol{\beta}) \boldsymbol{\psi}) (\mathbf{Y}_t \boldsymbol{\psi} - \boldsymbol{\xi}_t(\boldsymbol{\beta}) \boldsymbol{\psi})^\top \\ &= \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\psi} \boldsymbol{\psi}^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top \\ &= \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top. \end{aligned}$$

Ainsi, nous voyons que l'exécution de la régression non linéaire (9.47) ou (9.49) fournira exactement les mêmes estimations GNLS que la minimisation de la somme généralisée des résidus au carré (9.45).

Normalement, la matrice de covariance contemporaine  $\boldsymbol{\Sigma}$  ne sera pas connue et donc  $\boldsymbol{\psi}$  ne le sera pas également. Cependant, il est souvent aisé d'obtenir une estimation convergente de  $\boldsymbol{\Sigma}$ , disons  $\check{\boldsymbol{\Sigma}}$ . Pourvu que chaque équation individuelle, pour  $i = 1, \dots, m$ , soit identifiée (peut-être une hypothèse peu réaliste dans le cas de certains modèles multivariés non linéaires tels que les systèmes de demande), il est possible d'estimer chaque équation par OLS ou NLS afin d'obtenir la matrice de dimension  $n \times m$  des résidus  $\check{\mathbf{U}}$ . Alors, il est facile de voir que, sous des conditions assez générales

$$\check{\boldsymbol{\Sigma}} \equiv n^{-1} \check{\mathbf{U}}^\top \check{\mathbf{U}} \quad (9.50)$$

fournira une estimation convergente de  $\boldsymbol{\Sigma}$ . Etant donné  $\check{\boldsymbol{\Sigma}}$ , on peut facilement calculer  $\check{\boldsymbol{\psi}}$  en utilisant (9.46). Alors l'estimation NLS de (9.47), en remplaçant  $\boldsymbol{\psi}$  par  $\check{\boldsymbol{\psi}}$ , fournira les estimations par GNLS faisables qui, comme d'habitude,

sont asymptotiquement équivalentes aux estimations par GNLS ordinaires. Il s'agit de la procédure préconisée par Zellner (1962) dans le cas SUR.

Les conditions du premier ordre pour la minimisation de la somme généralisée des résidus au carré (9.45) peuvent être écrites de différentes façons. La raison fondamentale en est que la dérivée de  $\xi_{ti}(\boldsymbol{\beta})$  par rapport à  $\beta_j$ , le  $j^{\text{ième}}$  élément de  $\boldsymbol{\beta}$ , comprend nécessairement trois indices. Une approche consiste à définir  $\boldsymbol{\Xi}_t(\boldsymbol{\beta})$  comme une matrice de dimension  $k \times m$  avec comme élément type

$$\Xi_{t,ji}(\boldsymbol{\beta}) \equiv \frac{\partial \xi_{ti}(\boldsymbol{\beta})}{\partial \beta_j}.$$

Les conditions du premier ordre peuvent alors être écrites comme

$$\sum_{t=1}^n \boldsymbol{\Xi}_t(\tilde{\boldsymbol{\beta}}) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\tilde{\boldsymbol{\beta}}))^{\top} = \mathbf{0}. \quad (9.51)$$

Une seconde approche consiste à définir  $\mathbf{y}_i$  comme la  $i^{\text{ième}}$  colonne de  $\mathbf{Y}$  et  $\mathbf{x}_i(\boldsymbol{\beta})$  comme le vecteur des fonctions de régression pour la  $i^{\text{ième}}$  équation du système, c'est-à-dire un vecteur de dimension  $n$  avec comme élément type  $\xi_{ti}(\boldsymbol{\beta})$ . Alors, si on classe les dérivées de  $\mathbf{x}_i(\boldsymbol{\beta})$  par rapport à  $\boldsymbol{\beta}$  dans une matrice de dimension  $n \times k$   $\mathbf{Z}_i(\boldsymbol{\beta})$  avec l'élément type

$$(\mathbf{Z}_i)_{tj}(\boldsymbol{\beta}) \equiv \frac{\partial \xi_{ti}(\boldsymbol{\beta})}{\partial \beta_j} \quad (9.52)$$

et si on peut désigner par  $\sigma^{ij}$  le  $(i, j)^{\text{ième}}$  élément de  $\boldsymbol{\Sigma}^{-1}$ , un peu d'algèbre montrera que (9.51) devient

$$\sum_{i=1}^m \sum_{j=1}^m \sigma^{ij} \mathbf{Z}_i^{\top}(\boldsymbol{\beta}) (\mathbf{y}_j - \mathbf{x}_j(\boldsymbol{\beta})) = \mathbf{0}. \quad (9.53)$$

Un cas qui présente un intérêt particulier survient quand il n'existe pas de restrictions croisées dans le système. Le vecteur paramétrique complet peut alors être partitionné comme  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 : \dots : \boldsymbol{\beta}_m]$ , où les éléments du vecteur  $\boldsymbol{\beta}_i$  de dimension  $k_i$  sont les paramètres qui apparaissent seulement dans la  $i^{\text{ième}}$  équation. Il faut, bien sûr, que  $\sum_{i=1}^m k_i = k$ . Les matrices  $\mathbf{Z}_i$  peuvent contenir certains éléments nuls dans ce cas, parce que  $\xi_{ti}$  dépend seulement des éléments de  $\boldsymbol{\beta}_i$ . Il est commode de définir les matrices  $\bar{\mathbf{Z}}_i$  de dimension  $n \times k_i$  sans les éléments nuls; l'élément type sera  $(\bar{\mathbf{Z}}_i)_{tj} \equiv \partial \xi_{ti} / \partial (\beta_i)_j$  pour  $j = 1, \dots, k_i$ . Ceci permet de décomposer les conditions du premier ordre (9.53) équation par équation, afin d'obtenir

$$\sum_{j=1}^m \sigma^{ij} \bar{\mathbf{Z}}_i^{\top}(\boldsymbol{\beta}_i) (\mathbf{y}_j - \mathbf{x}_j(\boldsymbol{\beta}_j)) = \mathbf{0}, \quad i = 1, \dots, m. \quad (9.54)$$

Il est clair de (9.54) que si  $\Sigma$  est proportionnelle à une matrice identité, les conditions du premier ordre se réduisent à celles de NLS équation par équation quand il n'existera aucune restriction croisée. Ceci implique qu'il ne peut jamais y avoir de gain à exécuter l'estimation d'un système à moins que les corrélations contemporaines des aléas soient non nulles. Dans le contexte des GNLS faisables, il est extrêmement improbable que la matrice de covariance d'erreur estimée  $\check{\Sigma}$  de (9.50) sera proportionnelle à une matrice identité même si la véritable matrice  $\Sigma$  l'est. Dans ce cas, les estimations du système et les estimations équation par équation seront numériquement, mais pas asymptotiquement, différentes. Si  $\Sigma$  est proportionnelle à une matrice identité, alors  $\psi$  le sera également. Alors le système empilé (9.49) devient

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1(\beta_1) \\ \vdots \\ \mathbf{x}_m(\beta_m) \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{bmatrix}, \quad (9.55)$$

où le vecteur  $\mathbf{u}_i$  de dimension  $n$  représente le vecteur des aléas associés à la  $i^{\text{ième}}$  équation. Si le système empilé (9.55) était estimé par NLS, la somme des résidus au carré serait simplement

$$\sum_{i=1}^m (\mathbf{y}_i - \mathbf{x}_i(\beta_i))^{\top} (\mathbf{y}_i - \mathbf{x}_i(\beta_i)).$$

Comme les éléments de chaque  $\beta_i$  n'apparaissent que dans un seul terme de la somme sur  $i$ , cette somme est minimisée par la minimisation de chaque terme séparément par rapport aux paramètres dont elle dépend. Ainsi, l'estimation NLS de (9.55) correspond simplement à l'estimation NLS équation par équation.

Dans le cas particulier du système linéaire sans aucune restriction croisée, les conditions du premier ordre (9.53) peuvent être directement utilisées afin d'obtenir des estimations par GLS ou par GLS faisables du vecteur paramétrique  $\beta$ . Ceci utilise la propriété que, comme nous l'avons vu dans la Section 9.3, n'importe quel estimateur GLS peut être interprété comme un estimateur IV simple pour un choix convenable des instruments. Dans ce cas, les fonctions de régression empilées pour le système peuvent être écrites comme

$$\mathbf{X}\beta \equiv \begin{bmatrix} \mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_m \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}.$$

Ici  $\mathbf{X}_i$  désigne la matrice de dimension  $n \times k_i$  des régresseurs qui apparaissent dans la  $i^{\text{ième}}$  équation du système. En termes de la notation de (9.54), nous avons  $\mathbf{X}_i = \bar{\mathbf{Z}}_i(\beta_i)$ , où  $\mathbf{X}_i$  ne dépend pas de  $\beta_i$  parce que le système est linéaire. Si nous supposons que la matrice de covariance contemporaine  $\Sigma$  est

connue, nous pouvons construire la matrice  $\mathbf{W}$  de dimension  $nm \times k$  comme

$$\mathbf{W} = \begin{bmatrix} \sigma^{11} \mathbf{X}_1 & \cdots & \sigma^{1m} \mathbf{X}_m \\ \vdots & \ddots & \vdots \\ \sigma^{m1} \mathbf{X}_1 & \cdots & \sigma^{mm} \mathbf{X}_m \end{bmatrix}. \quad (9.56)$$

Ainsi,  $\mathbf{W}$  est une matrice partitionnée avec un bloc type  $\sigma^{ij} \mathbf{X}_j$  de dimension  $n \times k_j$ . Si  $\Sigma$  n'est pas connue, mais peut être estimée, alors  $\Sigma$  devrait être remplacée dans (9.56) par  $\tilde{\Sigma}$ .

Il est facile de voir que l'estimateur GLS est le même que l'estimateur IV simple

$$\tilde{\beta} \equiv \begin{bmatrix} \tilde{\beta}_1 \\ \vdots \\ \tilde{\beta}_m \end{bmatrix} = (\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{y},$$

où  $\mathbf{y} \equiv [\mathbf{y}_1 \vdots \dots \vdots \mathbf{y}_m]$ . Cet estimateur, bien que donné explicitement par la formule précédente, peut être défini au moyen des conditions du premier ordre

$$\mathbf{W}^\top \mathbf{X} \tilde{\beta} = \mathbf{W}^\top \mathbf{y}.$$

Si on détaille ces conditions, en utilisant les définitions de  $\mathbf{X}$  et  $\mathbf{W}$ , on peut voir qu'elles sont identiques à (9.54). Ainsi, pour les SUR linéaires sans aucune restriction croisée, les estimations paramétriques GLS peuvent être obtenues en employant une procédure IV pour estimer la régression univariée empilée  $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ , dans laquelle la matrice  $\mathbf{W}$  définie dans (9.56) est utilisée comme matrice d'instruments. Naturellement, comme nous l'avons remarqué plus tôt, la matrice de covariance estimée sera incorrecte.

Nous avons vu il y a peu qu'il n'existe aucun gain asymptotique obtenu par l'estimation d'un ensemble de SUR comme un système d'équation par équation s'il n'existe aucune corrélation contemporaine des aléas associés aux différentes équations du système. Il existe un autre cas dans lequel l'estimation du système d'équation ne produit aucun gain, cette fois-ci parce que les deux méthodes d'estimation mènent à des estimations paramétriques identiques *numériquement*. Cela survient dans le contexte d'un SUR linéaire quand toutes les matrices de régresseurs  $\mathbf{X}_i$  dans (9.56) sont les mêmes. Les estimations paramétriques sont identiques parce que le Théorème de Kruskal (voir Section 9.3) s'applique.

Nous montrons ceci en démontrant que l'espace engendré par les instruments  $\mathbf{W}$  est le même que celui des régresseurs  $\mathbf{X}$  à chaque fois que  $\mathbf{X}_i = \mathbf{X}^*$ , disons, pour tout  $i = 1, \dots, m$ . Ainsi, comme cela apparaît clairement lorsque l'on interprète un estimateur GLS comme d'un estimateur IV,  $\mathbf{W}$  joue le rôle de  $\Omega^{-1} \mathbf{X}$  dans l'énoncé général du Théorème de Kruskal. L'espace engendré par les colonnes de  $\mathbf{W}$  est l'ensemble des vecteurs de dimension  $nm$  de la forme  $[\mathbf{X}^* \gamma_1 \vdots \dots \vdots \mathbf{X}^* \gamma_m]$ , pour des vecteurs arbitraires  $\gamma_i$  qui possèdent autant d'éléments que  $\mathbf{X}^*$  comporte de colonnes. Tous ces vecteurs de dimension

$nm$  de ce type peuvent aussi être générés comme des combinaisons linéaires des colonnes de  $\mathbf{X}$ , qui correspond simplement à une matrice bloc-diagonale formée de blocs identiques  $\mathbf{X}^*$  le long de la diagonale principale. Il s'ensuit que  $\mathcal{S}(\mathbf{W}) = \mathcal{S}(\mathbf{X})$ , et le résultat est prouvé.

On associe à chaque modèle de régression non linéaire multivariée une version particulière de la régression de Gauss-Newton. Pour la  $i^{\text{ième}}$  observation, cette régression peut être écrite comme

$$(\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))\boldsymbol{\psi} = \mathbf{b}^\top \boldsymbol{\Xi}_t(\boldsymbol{\beta})\boldsymbol{\psi} + \text{résidu.} \quad (9.57)$$

Dans la pratique cette régression sera exécutée sous la forme empilée. Définissons un ensemble de  $m$  matrices  $\mathbf{X}_i(\boldsymbol{\beta})$ , toutes de dimension  $n \times k$ , en termes des matrices  $\mathbf{Z}_i(\boldsymbol{\beta})$  introduites dans (9.52), comme suit:

$$\mathbf{X}_i(\boldsymbol{\beta}) = \sum_{j=1}^m \mathbf{Z}_j(\boldsymbol{\beta})\psi_{ji}.$$

Alors la GNR empilée est

$$\begin{bmatrix} (\mathbf{Y} - \boldsymbol{\xi}(\boldsymbol{\beta}))\psi_1 \\ \vdots \\ (\mathbf{Y} - \boldsymbol{\xi}(\boldsymbol{\beta}))\psi_m \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1(\boldsymbol{\beta}) \\ \vdots \\ \mathbf{X}_m(\boldsymbol{\beta}) \end{bmatrix} \mathbf{b} + \text{résidus.} \quad (9.58)$$

Les estimations OLS de la GNR (9.58) seront définies par les conditions du premier ordre

$$\left( \sum_{i=1}^m \mathbf{X}_i^\top(\boldsymbol{\beta}) \mathbf{X}_i(\boldsymbol{\beta}) \right) \ddot{\mathbf{b}} = \sum_{i=1}^m \mathbf{X}_i^\top(\boldsymbol{\beta}) (\mathbf{Y} - \boldsymbol{\xi}(\boldsymbol{\beta}))\psi_i. \quad (9.59)$$

Quelques manipulations de (9.59) basées sur la définition des  $\mathbf{X}_i$  et de  $\boldsymbol{\psi}$  dévoilent que ceci est équivalent à

$$\sum_{i=1}^m \sum_{j=1}^m \sigma^{ij} \mathbf{Z}_i^\top(\boldsymbol{\beta}) (\mathbf{y}_j - \mathbf{x}_j(\boldsymbol{\beta}) - \mathbf{Z}_j(\boldsymbol{\beta})\mathbf{b}) = \mathbf{0}. \quad (9.60)$$

Ainsi, nous voyons que la régression (9.58) possède toutes les propriétés que nous sommes en droit d'attendre de la régression de Gauss-Newton. Si nous l'évaluons en  $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ , la régression n'aura aucun pouvoir explicatif, parce que (9.60) est satisfaite avec  $\mathbf{b} = \mathbf{0}$  en vertu des conditions du premier ordre (9.53). La matrice de covariance estimée de la régression (9.58) avec  $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$  sera

$$\tilde{s}^2 \left( \sum_{i=1}^m \sum_{j=1}^m \sigma^{ij} \tilde{\mathbf{Z}}_i^\top \tilde{\mathbf{Z}}_j^\top \right)^{-1}, \quad (9.61)$$

où  $\tilde{s}^2$  est l'estimation de la variance que la procédure informatique pour la régression générera, qui tendra évidemment asymptotiquement vers 1 si  $\Sigma$  est véritablement la matrice de covariance contemporaine de  $U_t$ . Si (9.61) est réécrite comme une somme de contributions des observations successives, le résultat est

$$\tilde{s}^2 \left( \sum_{t=1}^n \tilde{\Sigma}_t \Sigma^{-1} \tilde{\Sigma}_t^\top \right)^{-1},$$

de laquelle il est clair que (9.61) est en fait le véritable estimateur GNLS de la matrice de covariance.

Nous pouvons également exécuter la GNR empilée (9.58) avec toutes les quantités évaluées en un ensemble d'estimations ML  $\hat{\beta}$ , où les restrictions porte seulement sur  $\beta$  et non sur les éléments de  $\Sigma$ . La somme des carrés expliquée de cette régression sera

$$\left( \sum_{t=1}^n (Y_t - \xi_t) \Sigma^{-1} \dot{\Sigma}_t^\top \right) \left( \sum_{t=1}^n \dot{\Sigma}_t \Sigma^{-1} \dot{\Sigma}_t^\top \right)^{-1} \left( \sum_{t=1}^n (Y_t - \xi_t) \Sigma^{-1} \dot{\Sigma}_t^\top \right)^\top.$$

Ceci est à l'évidence une statistique LM. Elle peut être utilisée pour tester toutes sortes de restrictions sur  $\beta$ , et parmi celles-ci l'hypothèse que les aléas sont non corrélés en série. Pour en savoir plus sur les statistiques LM dans les modèles de régression multivariée, consulter Engle (1982a) et Godfrey (1988).

Les résultats antérieurs auraient pu être anticipés en vertu du fait qu'un modèle de régression multivariée peut toujours s'écrire comme un modèle de régression univariée. Néanmoins, il est utile d'avoir des résultats spécifiques pour les modèles multivariés. En particulier, la possibilité de calculer les régressions de Gauss-Newton fournit une façon commode d'obtenir les estimations par GNLS, pour vérifier que ces estimations sont précises, pour calculer les estimations de la matrice de covariance, et pour calculer les statistiques de test LM pour les restrictions sur  $\beta$ . Evidemment, tous ces résultats restent également valables pour les GNLS faisables, où  $\Sigma$  n'est pas disponible mais où l'estimation convergente  $\tilde{\Sigma}$  l'est.

## 9.9 L'ESTIMATION ML DES RÉGRESSIONS MULTIVARIÉES

Le principal concurrent des GLS faisables est l'estimation par maximum de vraisemblance basée sur l'hypothèse d'une distribution normale des aléas. Comme nous l'avons vu dans la Section 9.6, les estimations ML seront convergentes même si cette hypothèse est fausse, et c'est ce qui lui confère son statut d'hypothèse raisonnable. Ainsi, le modèle est maintenant

$$Y_t = \xi_t(\beta) + U_t, \quad U_t \sim \text{NID}(\mathbf{0}, \Sigma).$$

La densité de  $U_t$  est

$$(2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} U_t \Sigma^{-1} U_t^\top\right).$$

par conséquent, celle de  $\mathbf{Y}_t$  est

$$(2\pi)^{-m/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top\right).$$

De là, la fonction de logvraisemblance  $\ell(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$  est

$$-\frac{mn}{2} \log(2\pi) - \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top. \quad (9.62)$$

Notons qu'ici le dernier terme correspond précisément à l'opposé de la moitié de la somme généralisée des résidus aux carrés (9.45). Ainsi, si  $\boldsymbol{\Sigma}$  était connue, les estimations ML de  $\boldsymbol{\beta}$  seraient identiques aux estimations GLS.

La première étape dans la maximisation de  $\ell(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$  consiste à la concentrer par rapport à  $\boldsymbol{\Sigma}$ . Puisque  $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}^{-1}|^{-1}$ , (9.62) peut s'exprimer uniquement en termes de la matrice inverse  $\boldsymbol{\Sigma}^{-1}$ , de telle sorte qu'il est plus facile de concentrer la logvraisemblance en utilisant les conditions du premier ordre données en la dérivant par rapport aux éléments de  $\boldsymbol{\Sigma}^{-1}$ . La matrice des dérivées partielles ainsi obtenue est (consulter l'Annexe A pour les détails de la dérivation)

$$\frac{\partial \ell}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{n}{2} \boldsymbol{\Sigma} - \frac{1}{2} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})). \quad (9.63)$$

Le fait de poser que l'expression de droite dans (9.63) est égale à zéro donne

$$-\frac{n}{2} \boldsymbol{\Sigma} = -\frac{1}{2} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})),$$

d'où nous voyons que

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})). \quad (9.64)$$

Ainsi, l'estimateur ML de  $\boldsymbol{\Sigma}$  est exactement ce à quoi l'on pourrait s'attendre, c'est-à-dire la matrice des sommes des carrés et des produits croisés des résidus, divisée par la taille de l'échantillon.

Nous pouvons facilement substituer (9.64) dans le dernier terme de (9.62) si nous observons que la trace d'un scalaire est précisément le scalaire lui-même et que la trace d'un produit matriciel est invariante à une permutation cyclique des facteurs du produit. Nous obtenons

$$\begin{aligned} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top &= \text{Tr}\left((\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top\right) \\ &= \text{Tr}\left(\boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))\right). \end{aligned}$$

La somme sur  $t$  donne

$$\begin{aligned} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top &= \sum_{t=1}^n \text{Tr} \left( \boldsymbol{\Sigma}^{-1} (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \right) \\ &= \text{Tr} \left( \boldsymbol{\Sigma}^{-1} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \right) \\ &= \text{Tr} (\boldsymbol{\Sigma}^{-1} n \boldsymbol{\Sigma}) = mn. \end{aligned}$$

Ainsi, la fonction de logvraisemblance concentrée qui correspond à (9.62) est

$$\begin{aligned} \ell^c(\mathbf{Y}, \boldsymbol{\beta}) &= C - \frac{n}{2} \log \left| \frac{1}{n} \sum_{t=1}^n (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta}))^\top (\mathbf{Y}_t - \boldsymbol{\xi}_t(\boldsymbol{\beta})) \right| \\ &= C - \frac{n}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\beta})|, \end{aligned} \quad (9.65)$$

où  $\boldsymbol{\Sigma}(\boldsymbol{\beta})$  a été définie implicitement, et  $C$ , une constante qui ne dépend pas de  $\boldsymbol{\beta}$ , est égale à

$$- \frac{mn}{2} (\log(2\pi) + 1).$$

L'expression (9.65) est l'analogue multivarié de la fonction de logvraisemblance (8.82) pour des modèles de régression non linéaires univariée.

De (9.65), nous voyons que pour obtenir les estimations ML  $\hat{\boldsymbol{\beta}}$  nous devons minimiser le logarithme du déterminant de la matrice de covariance contemporaine,  $|\boldsymbol{\Sigma}(\boldsymbol{\beta})|$ . Ceci peut être fait très facilement en utilisant la règle de calcul des dérivées des logarithmes des déterminants donnée dans l'Annexe A. Cette règle stipule que si  $\mathbf{A}$  est une matrice de dimension  $m \times m$  non singulière, alors la dérivée de  $\log |\mathbf{A}|$  par rapport au  $(i, j)$ <sup>ième</sup> élément de  $\mathbf{A}$  est le  $(j, i)$ <sup>ième</sup> élément de  $\mathbf{A}^{-1}$ . Il vient que la dérivée de  $\log |\boldsymbol{\Sigma}(\boldsymbol{\beta})|$  par rapport à  $\beta_i$  est

$$\begin{aligned} \frac{\partial \log |\boldsymbol{\Sigma}(\boldsymbol{\beta})|}{\partial \beta_i} &= \sum_{j=1}^m \sum_{l=1}^m \frac{\partial \log |\boldsymbol{\Sigma}(\boldsymbol{\beta})|}{\partial \sigma_{jl}} \frac{\partial \sigma_{jl}(\boldsymbol{\beta})}{\partial \beta_i} \\ &= \sum_{j=1}^m \sum_{l=1}^m (\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}))_{lj} \frac{\partial \sigma_{jl}(\boldsymbol{\beta})}{\partial \beta_i} \\ &= \text{Tr} \left( \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}) \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\beta})}{\partial \beta_i} \right). \end{aligned}$$

Il est facile de voir que

$$\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\beta})}{\partial \beta_i} = - \frac{2}{n} \sum_{t=1}^n \mathbf{U}_t^\top(\boldsymbol{\beta}) \frac{\partial \boldsymbol{\xi}_t(\boldsymbol{\beta})}{\partial \beta_i},$$



d'où l'on peut voir que le gradient de (9.65) est

$$\sum_{t=1}^n \Xi_t(\beta) \Sigma(\beta)^{-1} (\mathbf{Y}_t - \xi_t(\beta))^\top. \quad (9.66)$$

En posant le gradient égal à zéro, nous retrouvons les conditions du premier ordre (9.51) obtenues de la méthode des GNLS, mais avec  $\Sigma(\beta)$  comme matrice de covariance.

Dans le cas des modèles de régression univariée, le fait que les estimations par moindres carrés soient choisies de façon à minimiser la somme des résidus au carré assure que, en moyenne, les résidus seront plus petits que les véritables aléas. Pour la même raison, le fait que les estimations ML minimisent le déterminant de la matrice de covariance contemporaine du modèle assure que, en moyenne, les résidus associés à ces estimations seront à la fois trop petits et trop fortement corrélés les uns aux autres. Nous observons les deux effets, parce que le déterminant de la matrice de covariance peut être construit plus petit soit en réduisant les sommes des résidus au carré associées aux équations individuelles soit en augmentant la corrélation entre les différentes équations. Ceci est probablement d'un intérêt plus appréciable lorsque  $m$  et/ou  $k$  sont grands relativement à  $n$ .

Il est intéressant de considérer la matrice d'information pour le modèle (9.43). Comme pour tous les modèles de régression, la matrice d'information se révélera être bloc-diagonale entre le bloc qui correspond à  $\beta$  et celui qui correspond à  $\Sigma$  ou, de manière équivalente, à  $\Sigma^{-1}$ . Pour constater ceci, observons de (9.63) que

$$\frac{\partial \ell_t}{\partial \Sigma^{-1}} = \frac{1}{2} \Sigma - \frac{1}{2} (\mathbf{Y}_t - \xi_t(\beta))^\top (\mathbf{Y}_t - \xi_t(\beta)).$$

Ceci est une matrice symétrique de dimension  $m \times m$  à  $m(m+1)/2$  éléments indépendants. Un de ses éléments types est

$$\frac{\partial \ell_t}{\partial \sigma^{ij}} = \frac{1}{2} \sigma_{ij} - \frac{1}{2} (y_{ti} - \xi_{ti}(\beta))^\top (y_{tj} - \xi_{tj}(\beta)). \quad (9.67)$$

A partir de (9.66), nous voyons également que le gradient de  $\ell_t$  par rapport à  $\beta$  est

$$\Xi_t(\beta) \Sigma^{-1} (\mathbf{Y}_t - \xi_t(\beta))^\top. \quad (9.68)$$

Si nous multiplions (9.67) par (9.68), le produit impliquera, selon le choix de  $i$  et de  $j$ , soit une soit trois occurrences de chaque composante de  $\mathbf{Y}_t - \xi_t(\beta) = \mathbf{U}_t$ . Parce que les premier et troisième moments des aléas sont nuls (une conséquence de la normalité), un tel produit doit avoir une espérance nulle. Ainsi, la matrice d'information doit être bloc-diagonale entre  $\beta$  et  $\Sigma$ .

Considérons à présent le bloc  $(\beta, \beta)$  de la matrice d'information. Par définition, il s'agit de la limite de l'espérance de  $1/n$  fois le produit extérieur du gradient, à savoir

$$\begin{aligned} \mathcal{I}_{\beta\beta} &= \lim_{n \rightarrow \infty} E \left( \frac{1}{n} \sum_{t=1}^n \Xi_t(\beta) \Sigma^{-1} (\mathbf{Y}_t - \xi_t(\beta))^\top (\mathbf{Y}_t - \xi_t(\beta)) \Sigma^{-1} \Xi_t^\top(\beta) \right) \\ &= \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \Xi_t(\beta) \Sigma^{-1} \Sigma \Sigma^{-1} \Xi_t^\top(\beta) \right) \\ &= \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \Xi_t(\beta) \Sigma^{-1} \Xi_t^\top(\beta) \right). \end{aligned}$$

Ainsi, nous concluons que

$$n^{1/2}(\hat{\beta} - \beta_0) \stackrel{a}{\sim} N \left( \mathbf{0}, \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \Xi_t(\beta) \Sigma^{-1} \Xi_t^\top(\beta) \right)^{-1} \right). \quad (9.69)$$

Notons que, excepté pour les facteurs  $\tilde{s}^2$ , l'estimation de la matrice de covariance (9.61) obtenue en exécutant la régression de Gauss-Newton est précisément l'estimation que le résultat (9.69) suggérerait d'utiliser. Si la GNR est calculée aux estimations ML  $\hat{\beta}$ , la variance d'erreur estimée pour cette régression,  $\hat{s}^2$ , sera égale à

$$\begin{aligned} & \frac{1}{mn - k} \sum_{t=1}^n (\mathbf{Y}_t - \hat{\xi}_t) \hat{\psi} \hat{\psi}^\top (\mathbf{Y}_t - \hat{\xi}_t)^\top \\ &= \frac{1}{mn - k} \sum_{t=1}^n (\mathbf{Y}_t - \hat{\xi}_t) \hat{\Sigma}^{-1} (\mathbf{Y}_t - \hat{\xi}_t)^\top = \frac{mn}{mn - k}. \end{aligned} \quad (9.70)$$

Ici, la dernière égalité provient d'un argument presque identique à celui utilisé pour établir (9.65). Comme il est évident que (9.70) tend asymptotiquement vers 1, l'expression (9.61), qui est dans ce cas

$$\frac{mn}{mn - k} \left( \sum_{t=1}^n \hat{\Xi}_t \hat{\Sigma}^{-1} \hat{\Xi}_t^\top \right)^{-1},$$

fournit une manière naturelle et très commode d'estimer la matrice de covariance de  $\hat{\beta}$ .

Maintenant nous avons établi tous les principaux résultats intéressants concernant l'estimation des modèles de régression multivariée non linéaire. Puisque tous ces résultats ont été établis en termes de modèles généraux et abstraits, il peut être utile de les rendre plus concrets si nous indiquons précisément comment notre notation générale se relie au cas du système de

dépense linéaire dont nous avons discuté plus tôt. Pour être concret, nous supposons que  $m = 2$ , ce qui signifie qu'il y a en tout trois marchandises. Alors nous voyons que

$$\mathbf{Y}_t = [s_{t1} \quad s_{t2}];$$

$$\boldsymbol{\beta} = [\alpha_1 \vdots \alpha_2 \vdots \gamma_1 \vdots \gamma_2 \vdots \gamma_3];$$

$$\boldsymbol{\xi}_t(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\gamma_1 p_{1t}}{E_t} + \frac{\alpha_1}{E_t} \left( E_t - \sum_{j=1}^3 p_{jt} \gamma_j \right) & \frac{\gamma_2 p_{2t}}{E_t} + \frac{\alpha_2}{E_t} \left( E_t - \sum_{j=1}^3 p_{jt} \gamma_j \right) \end{bmatrix};$$

$$\boldsymbol{\Xi}_t(\boldsymbol{\beta}) = \begin{bmatrix} \left( E_t - \sum_{j=1}^3 p_{jt} \gamma_j \right) / E_t & 0 \\ 0 & E_t - \sum_{j=1}^3 p_{jt} \gamma_j \\ (1 - \alpha_1) p_{1t} / E_t & -\alpha_2 p_{1t} / E_t \\ -\alpha_1 p_{2t} / E_t & (1 - \alpha_2) p_{2t} / E_t \\ -\alpha_1 p_{3t} / E_t & -\alpha_2 p_{3t} / E_t \end{bmatrix}.$$

Etablir la GNR pour tester l'hypothèse que  $\gamma_1 = \gamma_2 = \gamma_3 = 0$ , où les estimations soumises à cette restriction ont été obtenues, peut être un exercice utile.

Notre traitement des modèles multivariés a été relativement bref. Un traitement plus complet, mais seulement pour les modèles SUR, peut être trouvé chez Srivastava et Giles (1987), qui est également une source excellente pour les références concernant la littérature économétrique et statistique sur ce sujet.

## 9.10 MODÉLISATION DES DONNÉES À DEUX DIMENSIONS

De nombreux ensembles de données comportent à la fois une dimension temporelle et une autre dimension, dite transversale. Par exemple, elles peuvent couvrir 40 années de données sur 20 pays, ou 132 trimestres de données sur 50 états. L'avantage de tels ensembles de données est que la taille d'échantillonnage est habituellement assez grande (pour les exemples ci-dessus,  $40 \times 20 = 800$  et  $132 \times 50 = 6600$ ), ce qui signifie qu'ils devraient être potentiellement très porteurs d'information concernant les paramètres à estimer. L'inconvénient est qu'il est nécessaire de prendre en compte la nature bidimensionnelle des données. Un type particulier de données à deux dimensions survient quand le même échantillon d'individus, de ménages, ou de firmes est observé à deux ou plusieurs reprises dans le temps. Les données de ce type sont souvent appelées **données de panel**. Un ensemble de données de panel se compose généralement d'un assez petit nombre d'observations temporelles sur un grand nombre d'unités de la dimension transversale. Le déséquilibre entre les deux dimensions de l'échantillon peut rendre nécessaire

d'utiliser des techniques spéciales, et peut infirmer la théorie asymptotique standard.

Si nous indexons par  $t$  la dimension temporelle des données et par  $i$  la dimension transversale, nous pouvons écrire un modèle de régression univariée non linéaire pour les données à deux dimensions comme

$$y_{ti} = x_{ti}(\beta) + u_{ti}, \quad t = 1, \dots, T, \quad i = 1, \dots, n. \quad (9.71)$$

Il y a  $T$  périodes de temps et  $n$  groupes en coupe transversale, pour un total de  $nT$  observations. Si nous voulions supposer que les  $u_{ti}$  sont homoscédastiques et indépendants, nous pourrions simplement estimer (9.71) par NLS. Mais souvent cela ne sera pas une hypothèse réaliste. La variance de  $u_{ti}$  pourrait bien varier systématiquement avec  $t$  ou  $i$  ou les deux à la fois. De plus, il semble plausible que les aléas  $u_{ti}$  et  $u_{tj}$  seront corrélés pour un quelconque  $i \neq j$  si certains chocs affectent plusieurs groupes de la dimension transversale au même instant. De façon similaire, il semble plausible que les aléas  $u_{ti}$  et  $u_{si}$  soient corrélés pour un quelconque  $t \neq s$  si certains chocs affectent le même groupe en plus d'un instant du temps. Il est difficile de dire a priori si un quelconque manquement à cette hypothèse i.i.d. surviendra pour n'importe quel ensemble de données. Mais si c'est le cas, et que nous appliquons simplement NLS, nous obtiendrons une matrice de covariance estimée qui sera non convergente et pourra conduire à de sérieuses erreurs d'inférence. Dans certaines circonstances, nous pouvons même obtenir des estimations paramétriques non convergentes.

En principe, la gestion des manquements à l'hypothèse i.i.d. des types que nous venons de décrire est assez directe. On écrit simplement la matrice de covariance supposée de  $u_{ti}$  comme une fonction d'un ou de plusieurs paramètres inconnus, on utilise les moindres carrés pour obtenir des résidus à partir desquels on estime ces paramètres de manière convergente, et on applique ensuite les GLS faisables. De manière alternative, on peut utiliser le maximum de vraisemblance pour estimer simultanément les paramètres de la fonction de régression et les ceux de la matrice de covariance. En pratique, naturellement, il n'est pas toujours facile d'appliquer cette méthode, et il existe une littérature importante sur les techniques particulières pour procéder ainsi. Chamberlain (1984), Hsiao (1986), Judge, Hill, Griffiths, Lütkepohl, et Lee (1985, Chapitre 13), et Greene (1990a, Chapitre 16) constituent des références utiles. Dans cette section, nous ne discuterons que d'un petit nombre des techniques les plus simples et les plus largement applicables pour traiter des données à deux dimensions.

Quand soit  $T$  soit  $n$  est assez petit mais l'autre est raisonnablement grand, il est naturel de remanier le modèle univarié (9.71) en un modèle multivarié. Supposons, pour être concret, qu'il n'existe que quelques unités de la dimension transversale et de nombreuses périodes temporelles. Alors il semble naturel de grouper les observations allant de  $t1$  à  $tn$  dans un vecteur  $\mathbf{u}_t$  et de

supposer que

$$\mathbf{u}_t \sim \text{IID}(\mathbf{0}, \Sigma).$$

Ainsi, nous supposons que  $u_{ti}$  est en général corrélé avec  $u_{tj}$  pour  $i \neq j$  et que  $\mathbf{u}_t$  n'est pas corrélé avec  $\mathbf{u}_s$  pour  $t \neq s$ . Avec cette spécification d'erreur, le modèle univarié (9.71) devient un cas particulier du modèle de régression non linéaire multivariée (9.40) et peut être estimé soit par GLS faisables (Section 9.8) soit par maximum de vraisemblance (Section 9.9). Naturellement, il existera de nombreuses restrictions croisées, car les paramètres dans toutes les équations sont supposés être les mêmes, mais l'une ou l'autre de ces techniques devrait être capable de les traiter sans difficulté.

Le traitement d'un modèle tel que (9.71) comme un modèle multivarié est attrayant parce que l'on peut employer des logiciels standards pour l'estimation de tels modèles. De plus, il devient naturel de tester l'hypothèse (pas toujours plausible) que la même fonction de régression  $x_{ti}(\beta)$  s'applique à toutes les unités de la dimension transversale. Une exigence minimale est d'être toujours capable de vérifier que chaque unité peut avoir une ordonnée à l'origine différente. Ceci peut être fait de différentes manières. Deux possibilités existent, la première consistant à estimer le modèle non contraint et ensuite à calculer un test LR ou un test équivalent et la seconde consiste à calculer un test LM basé sur une GNR telle que (9.58). Par ailleurs, on voudrait sûrement pouvoir tester la corrélation des aléas à travers les périodes de temps. Ceci peut être réalisé en utilisant les tests standards pour la corrélation en série dans les modèles multivariés, qui peuvent également se baser sur la GNR (9.58). Ce thème sera abordé très brièvement dans la Section 10.11; consulter aussi Engle (1984) et Godfrey (1988). On peut également vouloir tester l'hétéroscédasticité à travers les périodes de temps, ce qui peut être réalisé par des extensions directes des techniques dont nous discuterons dans les Chapitres 11 et 16.

Bien que l'approche du traitement d'un modèle univarié estimé à l'aide de données à deux dimensions en un modèle multivarié possède de nombreux attraits, elle peut ne pas être pertinente si  $n$  et  $T$  sont tous deux assez importants. Par exemple, supposons que  $n = 30$  et  $T = 40$ . Alors un modèle multivarié qui traite chaque unité de la dimension transversale séparément aura 30 équations, et la matrice  $\Sigma$  comportera  $\frac{1}{2}(30 \times 31) = 465$  éléments distincts, qui devront être estimés individuellement avec seulement 40 observations. L'estimation d'un modèle à 30 équations est tout à fait réalisable. Cependant, avec seulement 40 observations, il sera difficile d'obtenir de bonnes estimations de  $\Sigma$ , et nous pourrions donc nous attendre à ce que les propriétés avec des échantillons finis des estimations GLS et ML soient pauvres.

Une seconde approche, très populaire, consiste à utiliser ce qui est appelé un **modèle à erreurs composées**. L'idée est de modéliser  $u_{ti}$  comme la composante de trois chocs individuels, chacun étant supposé être indépendant des autres:

$$u_{ti} = e_t + v_i + \varepsilon_{ti}. \quad (9.72)$$

Ici,  $e_t$  affecte toutes les observations de la période temporelle  $t$ ,  $v_i$  affecte toutes les observations effectuées sur l'unité  $i$ , et  $\varepsilon_{ti}$  affecte seulement l'observation  $ti$ . Dans les versions les plus répandues des modèles à erreurs composées, les  $e_t$  sont supposés être indépendants à travers le temps  $t$ , les  $v_i$  sont supposés être indépendants à travers les unités  $i$ , et les  $\varepsilon_{ti}$  sont supposés être indépendants à travers le temps  $t$  et les unités  $i$ . Ces hypothèses peuvent naturellement être relâchées, comme l'ont fait Revankar (1979) et Baltagi et Li (1991), mais nous n'en discuterons pas ici.

Il existe deux manières d'estimer un modèle de régression avec des aléas qui sont supposés être composés comme dans (9.72). La première consiste à estimer ce que l'on appelle un **modèle à effets fixes**, et la seconde approche consiste à estimer ce que l'on appelle un **modèle à effets aléatoires**. Ces deux approches sont conceptuellement très différentes. Dans la première nous estimons le modèle conditionnellement aux erreurs  $e_t$  et  $v_i$ , alors que dans la seconde, nous estimons le modèle de façon non conditionnelle. Un modèle à effets fixes peut être estimé par moindres carrés ordinaires (ou non linéaires), tandis qu'un modèle à effets aléatoires nécessite l'utilisation des GLS ou du ML. Un avantage du modèle à effets fixes est que, comme nous travaillons conditionnellement à  $e_t$  et  $v_i$ , nous n'avons pas besoin de supposer qu'ils sont indépendants des régresseurs. Cependant, comme nous le verrons, le modèle à effets aléatoires fournira des estimations plus efficaces lorsqu'il est approprié. Mundlak (1978) constitue une référence classique sur la relation entre les modèles à effets fixes et les modèles à effets aléatoires.

Pour faire simple tout en restant concret, nous supposerons dans la suite de cette section qu'il n'existe aucun choc temporel, ce qui implique que  $e_t = 0$  pour tout  $t$ . Ceci simplifie l'algèbre sans modifier la nature des résultats. Nous supposerons également que la fonction de régression pour l'observation  $ti$  est  $\mathbf{X}_{ti}\boldsymbol{\beta}$ . Sous ces hypothèses, le modèle à erreurs composées peut être écrit comme

$$y_{ti} = \mathbf{X}_{ti}\boldsymbol{\beta} + v_i + \varepsilon_{ti}. \quad (9.73)$$

L'idée du modèle à effets fixes consiste à traiter les  $v_i$  comme des paramètres inconnus et à les estimer conjointement avec  $\boldsymbol{\beta}$ . Ceci peut être fait en ajoutant  $n$  variables muettes  $D_{ti}^j$  à la régression (9.73), chacune étant égale à l'unité quand  $i = j$  et égale à zéro sinon. Naturellement, si  $\mathbf{X}_{ti}$  comprend un terme constant ou l'équivalent d'un terme constant, une des variables muettes devra être omise.

Dans la notation matricielle, la version à effets fixes de (9.73) peut être écrite comme

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{v} + \boldsymbol{\varepsilon}, \quad (9.74)$$

où  $\mathbf{v}$  est un vecteur de dimension  $n$  avec comme élément type  $v_i$ . A condition que les  $\varepsilon_{ti}$  soient i.i.d., le modèle (9.74) peut être estimé par OLS. En utilisant le Théorème FWL, nous voyons que l'estimateur des effets fixes de  $\boldsymbol{\beta}$  est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{M}_D \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_D \mathbf{y}, \quad (9.75)$$

où la matrice  $M_D$  est simplement la matrice qui calcule des écarts par rapport aux moyennes  $\bar{X}_i$  pour  $i = 1, \dots, n$ . Ainsi, un élément type de  $M_D X$  est

$$(M_D X)_{ti} = X_{ti} - \bar{X}_i.$$

Cette manipulation permet de calculer facilement  $\hat{\beta}$  même lorsque  $n$  prend des valeurs telles qu'il serait impossible d'exécuter la régression (9.74). Il suffit simplement de calculer les moyennes des groupes  $y_{.i}$  et  $X_{.i}$  pour tout  $i$  et de régresser  $y_{ti} - \bar{y}_{.i}$  sur  $X_{ti} - \bar{X}_i$  pour tout  $t$  et  $i$ . La matrice de covariance estimée devrait alors être ajustée pour tenir compte du fait que le nombre de degrés de liberté utilisé dans l'estimation est en fait  $n + k$  plutôt que  $k$ .

Parce que l'estimateur des effets fixes (9.75) dépend seulement des écarts de la régressande et des régresseurs par rapport à leurs moyennes de groupe respectives, il est parfois appelé l'**estimateur intra-groupes**. Comme le nom l'implique, il n'exploite pas le fait que les moyennes du groupe soient en général différentes pour des groupes différents. Cette propriété de l'estimateur peut être un avantage ou un inconvénient, selon les circonstances. Comme nous l'avons mentionné auparavant, il se peut que les effets transversaux  $v_i$  soient corrélés avec les régresseurs  $X_{ti}$  et par conséquent aussi avec les moyennes de groupe des régresseurs. Dans cette éventualité, l'estimateur OLS (sans effet fixe) basé sur l'échantillon complet serait non convergent, mais l'estimateur intra-groupes restera convergent. Cependant, si, par opposition, les effets fixes *sont* indépendants des régresseurs, l'estimateur intra-groupes n'est pas complètement efficace. Dans le cas extrême où une variable indépendante ne varie à l'intérieur des groupes, mais seulement entre les groupes, le coefficient correspondant à cette variable ne sera même pas identifiable par l'estimateur intra-groupes.

Un autre estimateur non efficace qui exploite seulement la variation sur les moyennes des groupes est appelé l'**estimateur inter-groupes**. Il peut être écrit comme

$$\hat{\beta} = (X^\top P_D X)^{-1} X^\top P_D y. \quad (9.76)$$

Comme  $P_D X_{ti} = \bar{X}_i$ , cet estimateur n'implique véritablement que  $n$  observations distinctes plutôt que  $nT$ . Il sera clairement non convergent si les effets transversaux, les  $v_i$ , sont corrélés avec les moyennes par groupe des régresseurs, les  $\bar{X}_i$ . L'estimateur OLS peut être écrit comme une moyenne pondérée (par des matrices) de l'estimateur intra-groupes et de l'estimateur inter-groupes:

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top y \\ &= (X^\top X)^{-1} (X^\top M_D y + X^\top P_D y) \\ &= (X^\top X)^{-1} X^\top M_D X \hat{\beta} + (X^\top X)^{-1} X^\top P_D X \hat{\beta}. \end{aligned}$$

Ainsi, nous voyons immédiatement que l'estimation par OLS sera non convergente toutes les fois que l'estimateur inter-groupes (9.76) est non convergent.

Même quand elle est convergente, l'estimation par OLS sera habituellement non efficace. Si les effets transversaux sont non corrélés avec les moyennes par groupe des régresseurs, alors nous voulons utiliser un modèle à effets aléatoires dans lequel les  $v_i$  ne sont pas traités comme fixes mais comme des composantes des aléas. Les OLS pondèrent toutes les observations de manière identique, mais ceci n'est pas optimal pour le modèle à erreurs composées (9.73). La variance de  $u_{ti}$  est, en utilisant une notation évidente,  $\sigma_v^2 + \sigma_\varepsilon^2$ . La covariance de  $u_{ti}$  avec  $u_{tj}$  est, par hypothèse, nulle pour  $i \neq j$ . Mais la covariance de  $u_{ti}$  avec  $u_{si}$  pour  $s \neq t$  est  $\sigma_v^2$ . Ainsi, si les données sont ordonnées en premier selon  $i$  et ensuite selon  $t$ , la matrice de covariance de  $u_{ti}$  peut être écrite comme

$$\begin{bmatrix} \Sigma & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma \end{bmatrix},$$

où  $\Sigma$  est la matrice de dimension  $T \times T$

$$\begin{bmatrix} \sigma_v^2 + \sigma_\varepsilon^2 & \sigma_v^2 & \cdots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + \sigma_\varepsilon^2 & \cdots & \sigma_v^2 \\ \vdots & \vdots & & \vdots \\ \sigma_v^2 & \sigma_v^2 & \cdots & \sigma_v^2 + \sigma_\varepsilon^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I} + \sigma_v^2 \boldsymbol{\iota} \boldsymbol{\iota}^\top.$$

Cette matrice de covariance illustre le fait que pour un  $i$  fixé, les erreurs sont équirellées; à comparer à (9.24).

Afin de calculer les estimations GLS, nous avons besoin de déterminer  $\Sigma^{-1/2}$ . Il est facile de vérifier que

$$\Sigma^{-1/2} = \frac{1}{\sigma_\varepsilon} (\mathbf{I} - \alpha \mathbf{P}_\iota),$$

où  $\mathbf{P}_\iota = T^{-1} \boldsymbol{\iota} \boldsymbol{\iota}^\top$  et  $\alpha$ , qui doit être compris entre 0 et 1, est défini par

$$\alpha = 1 - \frac{\sigma_\varepsilon}{(T\sigma_v^2 + \sigma_\varepsilon^2)^{1/2}}. \quad (9.77)$$

Ceci implique que l'élément type de  $\Sigma^{-1/2} \mathbf{y}_{.i}$  est  $\sigma_\varepsilon^{-1} (y_{ti} - \alpha \bar{y}_{.i})$ , et un élément type de  $\Sigma^{-1/2} \mathbf{X}_{.i}$  est  $\sigma_\varepsilon^{-1} (\mathbf{X}_{ti} - \alpha \bar{\mathbf{X}}_{.i})$ . Les estimations GLS peuvent alors être obtenues en exécutant la régression OLS

$$y_{ti} - \alpha \bar{y}_{.i} = (\mathbf{X}_{ti} - \alpha \bar{\mathbf{X}}_{.i}) \boldsymbol{\beta} + \text{résidu},$$

qui peut être écrite en termes matriciel comme

$$(\mathbf{I} - \alpha \mathbf{P}_D) \mathbf{y} = (\mathbf{I} - \alpha \mathbf{P}_D) \mathbf{X} \boldsymbol{\beta} + (\mathbf{I} - \alpha \mathbf{P}_D) \mathbf{u}. \quad (9.78)$$



En pratique, naturellement,  $\alpha$  sera inconnu, et il nous faudra employer les GLS faisables ou le maximum de vraisemblance. Cette première technique est très facile à mettre en œuvre, car nous pouvons obtenir les estimations des quantités dont nous avons besoin en estimant le modèle à effets fixes. Les aléas pour ce modèle sont simplement les  $\varepsilon_{ti}$ , et ainsi son estimation produira immédiatement une estimation convergente de  $\sigma_\varepsilon^2$ . Nous pouvons alors estimer  $\sigma_v^2$  de diverses manières, l'estimateur le plus simple étant la moyenne des estimations au carré des  $v_i$ . Cet estimateur sera aussi convergent, pourvu que  $T$  (et pas simplement  $nT$ ) puisse tendre vers l'infini. A l'aide de ces estimations de  $\sigma_\varepsilon^2$  et  $\sigma_v^2$ , nous pouvons facilement obtenir une estimation convergente de  $\alpha$  à partir de (9.77). Nous ne discuterons pas de l'estimation ML, qui est directe conceptuellement mais beaucoup plus difficile à calculer que les GLS faisables; la référence classique est Balestra et Nerlove (1966).

Il est intéressant de voir comment l'estimateur GLS défini par la régression (9.78) est relié à l'estimateur OLS et à l'estimateur intra-groupes (9.75). Lorsque  $\alpha = 0$ , l'estimateur GLS se confond évidemment avec l'estimateur OLS. Ceci a du sens parce que, à partir de (9.77), nous voyons que  $\alpha$  ne sera 0 que si  $\sigma_v = 0$ , auquel cas le terme d'erreur ne possède qu'un élément. Quand  $\alpha = 1$ , l'estimateur GLS se confond avec l'estimateur intra-groupes. Ceci a aussi du sens, parce que  $\alpha$  sera égal à 1 seulement si  $\sigma_\varepsilon = 0$ , auquel cas les aléas associés à la variation intra-groupes seront tous nuls. Ceci implique que nous pouvons obtenir des estimations parfaitement précises de  $\beta$  en utilisant l'estimateur intra-groupe. Dans chaque autre cas,  $\alpha$  sera compris entre 0 et 1, et l'estimateur GLS exploitera à la fois la variation intra-groupes et la variation inter-groupe.

Le problème avec les données de panel est que  $n$  est habituellement très grand et  $T$  est fréquemment très petit. Ainsi, les paramètres dont l'identification dépend de la variation des groupes transversaux sont normalement estimés de façon très satisfaisante, à la différence des paramètres dont l'identification dépend de la seule variation temporelle. On ne pourrait pas du tout s'attendre à estimer  $\sigma_v$  précisément dans un modèle à effets aléatoires, par exemple. Si on ne portait pas aucun intérêt à la variation temporelle, on utiliserait simplement un modèle à effets transversaux. Au lieu de soustraire explicitement les moyennes par groupes, nous pourrions calculer les différences premières de toutes les données par rapport à la dimension temporelle, de manière à faire disparaître les effets individuels. En pratique, cependant, nous sommes souvent intéressés par des paramètres qui ne sont pas identifiés seulement par la variation intra-groupes. Les économètres ont alors proposé un large éventail de procédures pour traiter des données de panel. Consulter, parmi tant d'autres, Hausman et Taylor (1981), Chamberlain (1984), Hsiao (1986), et Holtz-Eakin, Newey, et Rosen (1988).

## 9.11 CONCLUSION

Les GLS et GNLS sont des techniques d'estimation très importantes qui sont largement usitées en économétrie appliquée. Nous en rencontrerons des variantes dans la suite de cet ouvrage, plus particulièrement dans le Chapitre 10, où nous traitons de la corrélation en série, et dans le Chapitre 18, où nous traitons des techniques de systèmes complets pour estimer les modèles d'équations simultanées. Néanmoins, il est important de se souvenir que les GLS et les GNLS ne sont que des variantes masquées des moindres carrés. N'importe quelle erreur que l'on peut commettre en spécifiant un modèle estimé par OLS (telle que la spécification incorrecte de la fonction de régression ou une défaillance de la gestion de la corrélation en série ou de l'hétéroscédasticité) peut également être commise en spécifiant des modèles estimés par GLS, par GNLS, et par les méthodes variées du maximum de vraisemblance qui s'y rattachent. Il est alors tout aussi important de tester la mauvaise spécification de tels modèles que de tester le modèle de régression le plus simple. Les régressions de Gauss-Newton (9.14) et (9.58) fournissent souvent des manières commodes de le faire. Cependant, notre expérience nous révèle que le nombre de tests de spécification auquel un modèle est soumis est inversement relié à la difficulté d'estimation du modèle. Puisqu'il nous faut fournir habituellement un effort plus important pour estimer des modèles par GLS ou par GNLS et en particulier des modèles multivariés que pour estimer des modèles de régression univariée par OLS, les modèles estimés par GLS ou par GNLS sont souvent soumis à des tests de mauvaise spécification moins nombreux que ce que l'on imagine.

## TERMES ET CONCEPTS

données de panel	modèle de régression non linéaire
données à deux dimensions	multivariée
équivalence asymptotique des GNLS,	moindres carrés généralisés (GLS)
GNLS faisables, et ML	moindres carrés généralisés non
erreurs équadcorréliées	linéaires (GNLS)
estimateur GLS (Aitken)	moindres carrés pondérés
estimateur inter-groupes	régressions sans lien apparent
estimateur intra-groupes	(système SUR)
fonction d'utilité de Stone-Geary	restrictions croisées
fonction scédastique	somme généralisée des résidus au
GLS faisables et GNLS	carré
matrice de covariance contemporaine	système d'équation singulier
matrice de projection oblique	système de dépense linéaire
modèles à effets aléatoires	systèmes de demande
modèles à effets fixes	Théorème de Kruskal
modèles à erreurs composées	

# Chapitre 10

## Autocorrélation

### 10.1 INTRODUCTION

Le phénomène d'**autocorrélation**, pour lequel des résidus successifs apparaissent autocorrélés, est très fréquent dans les modèles estimés avec des données chronologiques. Par suite, les tests d'autocorrélation et l'estimation des modèles qui la prennent en compte sont des thèmes qui sont étudiés depuis très longtemps par les économètres, et la littérature est par conséquent vaste. Par bonheur, les résultats déjà établis au sujet des NLS, GNLS et du ML nous permettent de manipuler la plupart des problèmes associés à l'autocorrélation de manière assez directe.

Bien que les aléas puissent ne pas être indépendants dans n'importe quel genre de modèle, l'absence d'indépendance est la plupart du temps observée dans des modèles estimés avec des séries temporelles. En particulier, les observations qui sont proches dans le temps ont souvent des aléas qui apparaissent corrélés, alors que c'est rarement le cas de celles qui sont plus éloignées dans le temps. Nous disons qu'ils *apparaissent* corrélés parce qu'une mauvaise spécification de la fonction de régression peut conduire à une corrélation des résidus successifs, même si les véritables aléas ne le sont pas. Quoi qu'il en soit, que l'apparition de l'autocorrélation dans les modèles chronologiques soit affirmée ou pas, un modèle d'autocorrélation particulièrement simple a été largement adopté. Dans ce modèle, les aléas  $u_t$  sont supposés obéir au processus **autorégressif d'ordre un**, ou **AR(1)**,

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad |\rho| < 1. \quad (10.01)$$

Ce processus aléatoire indique que l'aléa au temps  $t$ ,  $u_t$ , est égal à une certaine fraction  $\rho$  de l'aléa au temps  $t - 1$  (avec un signe différent si  $\rho < 0$ ), plus un nouvel aléa ou **innovation**  $\varepsilon_t$  qui est homoscedastique et indépendant de toutes les innovations passées ou futures. Ainsi à chaque période, une partie de l'aléa correspond à l'aléa de la période précédente, quelque peu diminué et peut-être de signe différent et une partie correspond à l'innovation  $\varepsilon_t$ .

On appelle la condition  $|\rho| < 1$  **condition de stationnarité**. Elle garantit que la variance de  $u_t$  tend vers une valeur limite,  $\sigma^2$ , plutôt que de diverger

lorsque  $t$  augmente. En substituant successivement à  $u_{t-1}$ ,  $u_{t-2}$ ,  $u_{t-3}$ , et ainsi de suite dans (10.01), nous voyons que

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \dots$$

Ainsi, en utilisant la propriété d'indépendance des innovations  $\varepsilon_t, \varepsilon_{t-1} \dots$ , on voit que la variance de  $u_t$  est

$$\sigma^2 \equiv V(u_t) = \omega^2 + \rho^2\omega^2 + \rho^4\omega^2 + \rho^6\omega^2 + \dots = \frac{\omega^2}{1 - \rho^2}. \quad (10.02)$$

L'expression la plus à droite dans (10.02) n'est vraie que si la condition de stationnarité  $|\rho| < 1$  est vérifiée, puisque cette condition est nécessaire à la convergence de la somme infinie  $1 + \rho^2 + \rho^4 + \rho^6 + \dots$ . Dans les applications économétriques traditionnelles, où  $u_t$  est l'aléa joint à un modèle de régression, cette condition est pertinente, puisque nous ne voudrions certainement pas que la variance des aléas explose lorsque la taille de l'échantillon s'accroît.

Nous avons vu que, pour un processus stationnaire AR(1) qui s'est déroulé sur une période de temps conséquente, les aléas  $u_t$  auront tous une variance  $\sigma^2 = \omega^2/(1 - \rho^2)$ . Nous pouvons écrire

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \dots + \rho^{j-1}\varepsilon_{t-j+1} + \rho^j u_{t-j}, \quad (10.03)$$

exprimant la valeur de  $u_t$  comme une fonction de  $u_{t-j}$  et de toutes les innovations comprises entre les périodes  $t - j + 1$  et  $t$ . Par conséquent la covariance entre  $u_t$  et  $u_{t-j}$  peut se calculer comme

$$E((\varepsilon_t + \rho\varepsilon_{t-1} + \dots + \rho^{j-1}\varepsilon_{t-j+1} + \rho^j u_{t-j})u_{t-j}). \quad (10.04)$$

Puisque les innovations comprises entre les périodes  $t - j + 1$  et  $t$  sont indépendantes de  $u_{t-j}$ , la covariance (10.04) est simplement

$$E(\rho^j u_{t-j}^2) = \rho^j E(u_t^2) = \frac{\rho^j \omega^2}{1 - \rho^2} = \rho^j \sigma^2.$$

Nous en concluons que la matrice de covariance de  $\mathbf{u}$  est

$$\mathbf{\Omega} = \frac{\omega^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}, \quad (10.05)$$

où la matrice entre crochets est la matrice de corrélation de  $\mathbf{u}$ . Il est évident d'après (10.05) que chaque élément de  $\mathbf{u}$  est corrélé avec tous les autres éléments de  $\mathbf{u}$ , mais excepté lorsque  $|\rho|$  est très proche de 1, cette corrélation

tendra à disparaître assez rapidement au fur et à mesure que les périodes s'éloignent. Cela s'accorde bien à la fois à l'intuition et au comportement effectif des résidus de nombreux modèles de régression estimés à l'aide de séries temporelles. Ainsi il n'est pas étonnant que le processus AR(1) soit fréquemment utilisé dans les travaux économétriques appliqués.

Le processus AR(1) est bien sûr un cas très particulier. Il existe de nombreux autres processus aléatoires pouvant générer des aléas. Nous discuterons de certains d'entre eux dans les Sections 10.5 et 10.7. Cependant, parce que la plupart des problèmes soulevés par l'estimation et l'inférence dans les modèles d'autocorrélation sont déjà présents dans le cas AR(1), mais aussi parce que c'est le processus le plus fréquent en pratique, nous nous concentrerons pour l'instant sur le cas AR(1).

Le chapitre suit le plan suivant. Dans la section qui suit, nous discuterons des effets sur les estimations par moindres carrés d'une autocorrélation qui n'est pas prise en compte. Dans les deux sections suivantes, nous discuterons des méthodes pour l'estimation des modèles de régression qui permettent aux aléas d'être AR(1) mais qui ignorent la première observation. Puis, dans la Section 10.5, nous discuterons des processus AR d'ordre supérieur. La Section 10.6 traite des méthodes qui tiennent compte des observations initiales, et la Section 10.7 traite des aléas à moyenne mobile. Dans la Section 10.8, nous discuterons des tests d'autocorrélation et, dans la section qui suit, des tests des contraintes du facteur commun. La Section 10.10 traite de l'autocorrélation dans les modèles estimés par variables instrumentales. Finalement, dans la Section 10.11, nous discuterons brièvement de l'autocorrélation dans les modèles multivariés.

## 10.2 AUTOCORRÉLATION ET MOINDRES CARRÉS

Qu'advient-il si l'on utilise les moindres carrés pour estimer un modèle dans lequel les aléas sont en réalité autocorrélés? Par souci de simplicité, nous considérerons le cas linéaire, parce que tous les résultats se transposent au cas non linéaire de manière évidente. Supposons donc que nous estimions le modèle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I},$$

lorsque le processus générateur de données est en réalité

$$y_t = \mathbf{X}_t \boldsymbol{\beta}_0 + u_t, \quad u_t = \rho_0 u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega_0^2). \quad (10.06)$$

L'estimateur OLS est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

qui, sous le DGP (10.06), est égal à

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta}_0 + \mathbf{u}) = \boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}.$$

A condition que  $\mathbf{X}$  soit exogène,  $\hat{\beta}$  sera sans biais, parce que le fait que les  $u_t$  soient autocorrélés n'empêche pas  $E(\mathbf{X}^\top \mathbf{u})$  d'être nul. Si  $\mathbf{X}$  n'est pas exogène,  $\hat{\beta}$  sera convergent tant que  $\text{plim}(n^{-1} \mathbf{X}^\top \mathbf{u})$  est nulle.

Les inférences sur  $\beta$  ne seront cependant pas correctes. En supposant que  $\mathbf{X}$  est exogène, nous voyons que

$$\begin{aligned} E(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^\top &= E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Omega}_0 \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}, \end{aligned} \quad (10.07)$$

où  $\boldsymbol{\Omega}_0$  est la matrice  $\boldsymbol{\Omega}(\rho)$  définie en (10.05), évaluée en  $\rho_0$ . A l'évidence l'estimateur OLS  $s^2(\mathbf{X}^\top \mathbf{X})^{-1}$  de la matrice de covariance ne donnera pas une estimation convergente de (10.07). A l'exception de certains cas particuliers, il n'est pas possible de savoir si les estimations incorrectes des écarts types obtenues par OLS seront plus fortes ou plus faibles que les estimations correctes obtenues en prenant les racines carrées des éléments diagonaux de (10.07). Toutefois, l'analyse de cas particuliers suggère que pour des valeurs de  $\rho$  supérieures à 0 (la situation la plus fréquente), les écarts types OLS incorrects sont généralement trop faibles; consulter, entre autres, Nicholls et Pagan (1977), Sathe et Vinod (1974), et Vinod (1976).

L'expression (10.07) s'applique dans n'importe quel cas où les OLS sont employés à tort au lieu des GLS, et pas seulement dans les circonstances où les aléas obéissent à un processus AR(1). C'est également valable pour le résultat de l'absence de biais de  $\hat{\beta}$  lorsque  $\mathbf{X}$  est fixe et  $E(\mathbf{X}^\top \mathbf{u}) = \mathbf{0}$ . Mais souvenons-nous d'après la Section 9.5 que, même lorsque ces conditions sont satisfaites,  $\hat{\beta}$  peut ne pas converger si les aléas sont suffisamment autocorrélés. Nous pourrions donc conclure que lorsque les régresseurs sont fixes et la matrice de covariance des aléas telle qu'il n'y ait pas une autocorrélation trop importante entre les aléas, les estimations OLS seront convergentes, mais l'estimation OLS de la matrice de covariance ne le sera pas. On peut généralement calculer une estimation convergente de l'estimateur OLS de la matrice de covariance. Cependant, comme la démonstration du Théorème de Gauss-Markov dépendait de l'hypothèse  $E(\mathbf{u} \mathbf{u}^\top) = \sigma^2 \mathbf{I}$ , l'estimateur OLS n'est pas le meilleur estimateur linéaire sans biais lorsque cette hypothèse n'est pas retenue.

La discussion que nous venons de mener ignorait la présence de variables dépendantes retardées parmi les colonnes de  $\mathbf{X}$ . Lorsqu'au contraire on le suppose, les résultats varient considérablement, et l'estimateur OLS devient à la fois biaisé et non convergent. Le moyen le plus simple de s'en rendre compte est d'imaginer un élément de  $\mathbf{X}^\top \mathbf{u}$  correspondant à la variable dépendante retardée (ou à une des variables dépendantes retardées s'il en existe plusieurs dans  $\mathbf{X}$ ). Si la variable dépendante est retardée de  $j$  périodes, cet élément est

$$\sum_{t=1}^n y_{t-j} u_t. \quad (10.08)$$

Rappelons-nous maintenant l'expression (10.03), dans laquelle nous exprimons  $u_t$  comme une fonction de  $u_{t-j}$  et de toutes les innovations comprises entre les périodes  $t-j+1$  et  $t$ . Puisque  $y_{t-j}$  est égal à  $\mathbf{X}_{t-j}\boldsymbol{\beta} + u_{t-j}$ , il est clair d'après (10.03) que (10.08) ne peut pas être d'espérance nulle. Ainsi, nous en concluons que lorsque  $\mathbf{X}$  contient des variables dépendantes retardées et  $u_t$  est autocorrélé,

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{u} \right) \neq \mathbf{0}, \quad (10.09)$$

ce qui implique que

$$\text{plim}_{n \rightarrow \infty} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{u} \right) \neq \mathbf{0}. \quad (10.10)$$

Parce que  $\mathbf{X}^\top \mathbf{y}$  est prémultiplié par  $(\mathbf{X}^\top \mathbf{X})^{-1}$ , chaque élément de  $\mathbf{X}^\top \mathbf{u}$  affecte généralement  $\hat{\boldsymbol{\beta}}$ , à moins que la matrice  $\mathbf{X}^\top \mathbf{X}$  ne possède des caractéristiques très spéciales. Ainsi il est évident d'après (10.10) que chaque élément de  $\boldsymbol{\beta}$  sera estimé de façon non convergente, même s'il n'y a qu'une seule variable dépendante retardée et, par conséquent, un seul élément de (10.09) non nul.

La discussion précédente montre clairement pourquoi les économètres se sont tellement préoccupés de l'autocorrélation. Même lorsqu'il n'y a pas de variable dépendante retardée, ce phénomène rend les estimations par moindres carrés non efficaces et l'inférence qui se base sur elles peu valable. Lorsqu'il y a des variables dépendantes retardées, l'autocorrélation rend les estimations par moindres carrés biaisées et non convergentes. Néanmoins, il est important de se souvenir que de nombreux types de mauvaise spécification peuvent provoquer l'apparition d'autocorrélation. Ainsi la situation que nous venons d'analyser, dans laquelle le modèle était correctement spécifié à l'exception de la prise en compte de l'autocorrélation, ne rend probablement compte que d'une très faible proportion des cas dans lesquels les résidus d'un modèle de régression donnent l'apparence d'une autocorrélation.

### 10.3 ESTIMATION DES MODÈLES À ERREURS AR(1)

Supposons que l'on veuille estimer un modèle de régression non linéaire dont les aléas obéissent à un processus AR(1):

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.11)$$

Parce que  $u_{t-1} = y_{t-1} - x_{t-1}(\boldsymbol{\beta})$ , ce modèle peut être récrit comme

$$y_t = x_t(\boldsymbol{\beta}) + \rho(y_{t-1} - x_{t-1}(\boldsymbol{\beta})) + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.12)$$

qui est également un modèle de régression non linéaire, mais dont les aléas sont (par hypothèse) non autocorrélés. Puisque (10.12) est un modèle de régression

non linéaire dont les aléas ont de bonnes propriétés, il paraît naturel d'en faire une estimation non linéaire et de pratiquer des inférences en utilisant la régression de Gauss-Newton. La fonction de régression est simplement

$$x'_t(\beta, \rho) = x_t(\beta) + \rho(y_{t-1} - x_{t-1}(\beta)), \quad (10.13)$$

qui dépend autant de  $\rho$  que de  $\beta$ .

Il y a deux problèmes en puissance avec (10.12). En premier lieu, la fonction de régression  $x'_t(\beta, \rho)$  dépend nécessairement de  $y_{t-1}$ , que  $x_t(\beta)$  dépend ou non de n'importe quelle valeur de la variable dépendante. Comme nous l'avons vu dans le Chapitre 5, cette dépendance n'empêche pas les moindres carrés non linéaires d'avoir des propriétés asymptotiques attrayantes pourvu que certaines conditions de régularité soient satisfaites. On peut montrer que, tant que  $x_t(\beta)$  satisfait les conditions de régularité dictées par les Théorèmes 5.1 et 5.2 et que la condition de stationnarité  $|\rho| < 1$  est vérifiée, les moindres carrés non linéaires de (10.12) auront ces propriétés. Toutefois, si la condition de stationnarité n'était pas vérifiée, les résultats standards sur les moindres carrés non linéaires, et en particulier le Théorème 5.2, le théorème de normalité asymptotique, ne s'appliqueraient plus à (10.12).

En second lieu, comment gérer la première observation de (10.12)? Il est à croire que l'on ne dispose pas des valeurs pour  $y_0$  et pour toutes les variables prédéterminées et exogènes dont on a besoin pour évaluer  $x_0(\beta)$ , puisque si c'était le cas l'échantillon n'aurait pas débuté à l'observation correspondant à  $t = 1$ . Ainsi on ne peut pas évaluer  $x'_1(\beta, \rho)$ , qui dépend de  $y_0$  et de  $x_0(\beta)$ . La solution la plus simple pour résoudre ce problème est simplement d'omettre la première observation, ce qui nécessite que (10.12) soit valable pour les observations allant de 2 à  $n$  seulement. L'omission d'une observation ne modifie rien asymptotiquement, et nous pouvons sans danger en rejeter une lorsque la taille de l'échantillon est importante.

Une autre solution au problème de la gestion de la première observation consisterait à la traiter différemment de toutes les autres observations ultérieures, en définissant  $x'_1(\beta, \rho)$  comme  $x_1(\beta)$  au lieu de  $x_1(\beta) + \rho(y_0 - x_0(\beta))$ . Dans ce cas, l'aléa associé à l'observation 1 serait  $u_1$  plutôt que  $\varepsilon_1$ . Nous avons vu que, pourvu que le processus AR(1) soit stationnaire,  $u_t$  a une variance non conditionnelle égale à  $\omega^2/(1 - \rho^2)$  pour tout  $t$ , et également pour  $t = 1$ . En utilisant la première observation de cette façon, nous créerions de l'hétéroscédasticité: l'observation 1 aurait une variance égale à  $\omega^2/(1 - \rho^2)$ , alors que les observations restantes auraient toutes une variance de  $\omega^2$ . De plus, le paramètre  $\rho$  affecterait désormais non seulement la fonction de régression  $x'(\beta, \rho)$  mais aussi la variance de la première observation. Cela suggère que si nous voulons inclure la première observation, l'usage des moindres carrés non linéaires ne sera tout simplement pas approprié. En réalité, la prise en compte de la première observation compliquera substantiellement les choses, et nous discuterons donc de ce résultat plus en détail dans la Section 10.6.



Lorsque  $x_t(\boldsymbol{\beta})$  est non linéaire, l'estimation NLS de (10.12) nécessite évidemment un algorithme de maximisation non linéaire, tel que ceux basés sur la régression de Gauss-Newton présentés dans la Section 6.8. Dans la plupart des cas, cette estimation ne devrait pas être plus difficile à obtenir qu'une estimation du modèle correspondant où les aléas sont supposés être indépendants. Cette stratégie est également une stratégie adéquate lorsque le modèle est linéaire, c'est-à-dire lorsque  $x_t(\boldsymbol{\beta}) = \mathbf{X}_t\boldsymbol{\beta}$ . Dans la pratique cependant, la plupart des progiciels de régression proposent des procédures spécialisées pour estimer les modèles de régression linéaire avec des erreurs AR(1). Ces procédures peuvent parfois fonctionner plus efficacement que les moindres carrés non linéaires appliqués à (10.12), mais lorsqu'elles sont mises en œuvre par certains progiciels, elles peuvent conduire à des estimations de la matrice de covariance incorrectes dans certains cas (voir la Section 10.4).

Toutes les procédures d'estimation spécialisées pour les modèles de régression linéaire avec erreurs AR(1) utilisent le fait que, en fonction de la valeur de  $\rho$ , les estimations de  $\boldsymbol{\beta}$  peuvent facilement être obtenues par moindres carrés ordinaires. Pour le cas de la régression linéaire, on peut écrire (10.12) comme

$$y_t - \rho y_{t-1} = (\mathbf{X}_t - \rho \mathbf{X}_{t-1})\boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.14)$$

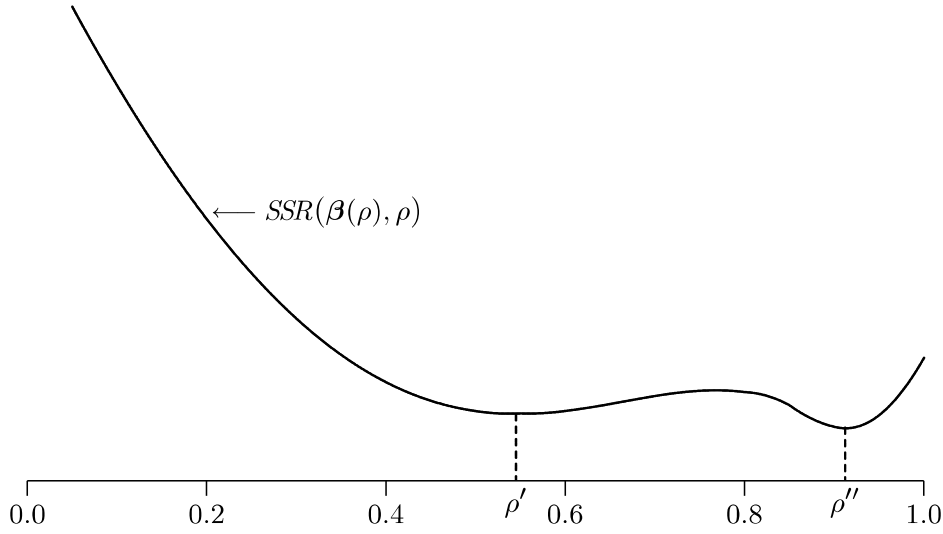
Ainsi, si on définit

$$y_t^*(\rho) \equiv y_t - \rho y_{t-1} \quad \text{et} \quad \mathbf{X}_t^*(\rho) \equiv \mathbf{X}_t - \rho \mathbf{X}_{t-1},$$

il suffit de régresser  $\mathbf{y}^*(\rho)$  sur  $\mathbf{X}^*(\rho)$  pour estimer  $\boldsymbol{\beta}$  en fonction de  $\rho$ .

Il est évident que toute estimation convergente de  $\rho$ , disons  $\hat{\rho}$ , entraînera une estimation convergente de  $\boldsymbol{\beta}$  lorsque  $\mathbf{y}^*(\hat{\rho})$  est régressé sur  $\mathbf{X}^*(\hat{\rho})$ . Il existe un grand nombre de techniques dans la littérature pour obtenir de telles estimations. Cependant, il est sans doute souhaitable pour des motifs d'efficacité de trouver effectivement l'estimation NLS  $\hat{\rho}$ , et compte tenu des facilités de calcul moderne il est difficile de justifier le contraire. Les deux techniques les plus largement utilisées pour trouver  $\hat{\rho}$  sont la **grille de valeurs** et la **recherche en alternance**. La première peut être utilisée dans de nombreuses autres situations pour lesquelles un modèle non linéaire est estimé aisément conditionnellement à un paramètre unique, et la seconde peut être utilisée dans un éventail large de situations pour lesquelles un modèle non linéaire est estimé aisément conditionnellement à chacun des deux sous-ensembles de paramètres qui ont une intersection vide. Ces deux techniques sont donc d'un certain intérêt.

L'usage de la grille de valeurs pour trouver les estimations  $(\hat{\boldsymbol{\beta}}, \hat{\rho})$  qui minimisent la somme des résidus au carré de (10.14), et par conséquent de (10.12) également, a été préconisée par Hildreth et Lu (1960), et on fait souvent référence à cette procédure dans la littérature en tant que **procédure de Hildreth-Lu**. L'idée de base est la simplicité même. Pour toute valeur de  $\rho$ ,



**Figure 10.1** Un cas où  $SSR(\beta(\rho), \rho)$  a deux minima

disons  $\rho^{(j)}$ , on peut exécuter une régression linéaire de  $\mathbf{y}^*(\rho^{(j)})$  sur  $\mathbf{X}^*(\rho^{(j)})$  pour trouver les estimations OLS

$$\beta^{(j)} = \left( \mathbf{X}^{*\top}(\rho^{(j)}) \mathbf{X}^*(\rho^{(j)}) \right)^{-1} \mathbf{X}^{*\top}(\rho^{(j)}) \mathbf{y}^*(\rho^{(j)}) \quad (10.15)$$

et une somme des carrés associée,  $SSR(\beta^{(j)}, \rho^{(j)})$ . Nous opérons pour des valeurs de  $\rho$  qui appartiennent à une grille de valeurs prédéterminée, c'est-à-dire toutes les valeurs comprises entre  $-0.999$  et  $0.999$  par intervalles d'amplitude  $0.1$  (c'est-à-dire,  $-0.999, -0.9, -0.8, \dots, 0.8, 0.9, 0.999$ ). Les valeurs d'arrêt sont ici  $\pm 0.999$  plutôt que  $\pm 1$  de manière à éviter l'infraction à la condition de stationnarité.

L'une des valeurs  $\rho^{(j)}$  de la grille, disons  $\rho^{(J)}$ , doit produire la plus faible valeur de  $SSR(\beta^{(j)}, \rho^{(j)})$ .<sup>1</sup> Puis, à condition que la grille soit assez précise, et en supposant que  $\rho^{(J)}$  n'est pas une valeur-borne de cette grille, il est raisonnable de s'attendre à ce que  $\hat{\rho}$  se situe quelque part entre  $\rho^{(J-1)}$  et  $\rho^{(J+1)}$ . Si  $\rho^{(J)}$  est l'une des valeurs-bornes, alors  $\hat{\rho}$  se situe probablement entre  $\rho^{(J)}$  et le point le plus proche de la grille. Dans les deux cas, on peut ensuite, soit établir une nouvelle grille sur cet intervalle plus court et opérer sur cette grille, soit débiter avec  $\rho^{(J)}$  une autre procédure de recherche, comme celle de la recherche en alternance décrite plus bas.

<sup>1</sup> Dans de rares cas, il peut y avoir deux, ou davantage,  $\rho^{(j)}$  qui entraînent des valeurs minimales de  $SSR(\beta^{(j)}, \rho^{(j)})$  identiques. Si ces  $\rho^{(j)}$  sont des valeurs assez proches, il n'y a pas de problème réel. Si ce n'est pas le cas, il faudrait approfondir l'étude de minima multiples en prenant une grille de valeurs plus fine au voisinage de chacune des  $\rho^{(j)}$ .

L'avantage de la grille de valeurs est qu'elle peut gérer des problèmes où il y a plus d'un minimum local. Considérons la Figure 10.1. Ici  $SSR(\beta(\rho), \rho)$  a deux minima locaux, l'un pour la valeur  $\rho'$  et l'autre qui correspond à un minimum global, pour la valeur  $\rho''$ . À condition que la grille soit suffisamment fine, la recherche sur la première grille conclura avec succès que  $\hat{\rho}$  est proche de 0.9. Au contraire, de nombreuses techniques d'estimation concluraient à tort que  $\hat{\rho}$  se situe en  $\rho'$ , en particulier si elles sont exécutées avec une valeur de départ  $\rho = 0$ . Cette propriété n'est pas seulement un avantage théorique en faveur de la grille de valeurs. Comme de nombreux auteurs l'ont montré—consulter Dufour, Gaudry, et Liem (1980) et Betancourt et Kelejian (1981)— $SSR(\beta(\rho), \rho)$  aura souvent des minima multiples. C'est particulièrement probable si  $\mathbf{X}$  contient une variable dépendante retardée. Dans ce cas, il y a souvent deux minima: l'un associé à une valeur faible de  $\rho$  et une valeur élevée du coefficient de la variable dépendante retardée, et l'autre associé à une forte valeur de  $\rho$  et une valeur faible de ce coefficient. L'un ou l'autre peut correspondre au minimum global.

La seconde procédure spécialisée pour les modèles de régression linéaire à aléas AR(1) dont nous allons parler est la recherche en alternance. Cette procédure, dont l'initiative revient à Cochrane et Orcutt (1949) pour le traitement de ce problème et à laquelle on fait généralement référence en tant que **procédure itérative de Cochrane-Orcutt**, est beaucoup plus largement utilisée que la grille de valeurs. Elle se base sur les faits que  $\beta$  est très facilement calculable en fonction de  $\rho$  et que  $\rho$  est aussi facilement calculable en fonction de  $\beta$ . L'algorithme de Cochrane-Orcutt débute avec une valeur initiale de  $\rho$ ,  $\rho^{(1)}$ , qui peut être égale à zéro ou qui peut être initialisée à une quelconque valeur différente si l'on dispose d'information a priori. Elle utilise ensuite cette valeur de  $\rho$  pour trouver une nouvelle valeur de  $\beta$ ,  $\beta^{(1)} \equiv \beta(\rho^{(1)})$ , qui est à son tour utilisée pour trouver une nouvelle valeur de  $\rho$ ,  $\rho^{(2)}$ , et ainsi de suite jusqu'à convergence. À chaque étape, la nouvelle valeur de  $\rho$  ou de  $\beta$  est celle qui minimise la somme des résidus au carré *conditionnellement* à une valeur donnée de  $\beta$  ou de  $\rho$ .

Pour toute valeur de  $\rho$ , disons  $\rho^{(j)}$ , une régression par OLS de  $\mathbf{y}^*(\rho^{(j)})$  sur  $\mathbf{X}^*(\rho^{(j)})$  entraîne  $\beta^{(j)} \equiv \beta(\rho^{(j)})$ , qui minimise  $SSR(\beta | \rho^{(j)})$ , c'est-à-dire la SSR en fonction de  $\rho^{(j)}$ ; la formule de  $\beta^{(j)}$  est donnée par (10.15). Étant donné  $\beta^{(j)}$ , on peut ensuite calculer les résidus

$$u_t^{(j)} \equiv y_t - \mathbf{X}_t \beta^{(j)}.$$

La SSR en tant que fonction de  $\rho$ , conditionnellement à  $\beta^{(j)}$ , est

$$SSR(\rho | \beta^{(j)}) = \sum_{t=2}^n \left( u_t^{(j)} - \rho u_{t-1}^{(j)} \right)^2. \quad (10.16)$$

Cette SSR est simplement celle de la régression linéaire de  $u_t^{(j)}$  sur  $u_{t-1}^{(j)}$  pour les observations allant de 2 à  $n$ , et par conséquent la valeur de  $\rho$  qui minimise

(10.16) est simplement l'estimation OLS de  $\rho$  de cette régression, qui est

$$\rho^{(j+1)} = \frac{\sum_{t=2}^n u_t^{(j)} u_{t-1}^{(j)}}{\sum_{t=2}^n (u_{t-1}^{(j)})^2}. \quad (10.17)$$

La procédure de Cochrane-Orcutt consiste donc en une succession de régressions par moindres carrés. La première consiste à régresser  $\mathbf{y}^*(\rho^{(1)})$  sur  $\mathbf{X}^*(\rho^{(1)})$ , la deuxième  $u_t^{(1)}$  sur  $u_{t-1}^{(1)}$ , la troisième  $\mathbf{y}^*(\rho^{(2)})$  sur  $\mathbf{X}^*(\rho^{(2)})$ , et ainsi de suite. A chaque étape,  $SSR(\boldsymbol{\beta}, \rho)$  est minimisée par rapport à  $\rho$  ou  $\boldsymbol{\beta}$ , et l'algorithme est autorisé à poursuivre tant qu'un certain critère de convergence n'est pas satisfait (habituellement on demande que  $\rho^{(j)}$  et  $\rho^{(j+1)}$  soient suffisamment proches). Une telle procédure doit en fin de compte converger vers un minimum local de la fonction SSR; voir Sargan (1964, Annexe A) et Oberhofer et Kmenta (1974). Hélas, rien ne garantit que ce minimum local soit également un minimum global. Il est par conséquent opportun de n'utiliser la procédure de Cochrane-Orcutt qu'après qu'une recherche préliminaire sur la grille de valeurs ait soit établi qu'il n'existe qu'un seul minimum local, soit situé approximativement le minimum global. Notons que bien que la procédure itérative de Cochrane-Orcutt fonctionne efficacement dans de nombreux cas, elle peut se révéler plus lente que l'usage d'un algorithme NLS basé sur la régression de Gauss-Newton.

Nous avons vu que la condition de stationnarité  $|\rho| < 1$  est essentielle pour que le processus AR(1) ait un sens et que les techniques d'estimation conventionnelles soient valables. Dans la pratique cependant, l'estimation NLS  $\hat{\rho}$  peut être supérieure à 1 en valeur absolue. Si cela survient, ou même si  $|\hat{\rho}|$  est très proche de 1, le praticien devrait sans doute considérer cela comme une évidente inadaptation du modèle. Peut-être le modèle devrait-il être spécifié à nouveau en termes des différences premières plutôt qu'en niveaux (Harvey, 1980), ou peut-être que la spécification de la fonction de régression, la spécification du processus AR(1) des aléas, ou les deux simultanément, est(sont) incompatible(s) avec les données. Nous discuterons d'une méthode de détection d'une mauvaise spécification dans les modèles qui semblent avoir des aléas AR(1) au cours de la Section 10.9.

La majeure partie de la discussion précédente s'est focalisée sur les méthodes pour obtenir des estimations NLS des modèles de régression linéaire où les aléas sont AR(1). Lorsque la taille de l'échantillon est très importante, il n'est pas utile de perdre du temps pour obtenir des estimations NLS, parce que les **estimations en une étape** qui sont asymptotiquement équivalentes aux NLS peuvent se révéler appropriées. Souvenons-nous d'après la Section 6.6 que, si  $\hat{\boldsymbol{\beta}}$  désigne n'importe quel vecteur d'estimations convergentes pour le modèle  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}$ , l'estimateur

$$\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}} + (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top (\mathbf{y} - \hat{\mathbf{x}}),$$

où  $\dot{\mathbf{X}} \equiv \mathbf{X}(\dot{\boldsymbol{\beta}})$  et  $\dot{\mathbf{x}} \equiv \mathbf{x}(\dot{\boldsymbol{\beta}})$ , est asymptotiquement équivalent à l'estimateur NLS  $\hat{\boldsymbol{\beta}}$ . Le terme que l'on ajoute ici à  $\hat{\boldsymbol{\beta}}$  est tout simplement l'estimation OLS de  $\mathbf{b}$  dans la régression de Gauss-Newton

$$\mathbf{y} - \dot{\mathbf{x}} = \dot{\mathbf{X}}\mathbf{b} + \text{résidus},$$

et la matrice de covariance OLS de cette dernière régression fournit une estimation valable de la matrice de covariance de  $\hat{\boldsymbol{\beta}}$ .

Avec l'omission de la première observation, comme d'habitude, on peut écrire un modèle de régression linéaire avec aléas AR(1) comme

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{y}_{-1} - \rho\mathbf{X}_{-1}\boldsymbol{\beta} + \varepsilon, \quad (10.18)$$

où  $y_{t-1}$  est l'élément type de  $\mathbf{y}_{-1}$  et  $\mathbf{X}_{t-1}$  est la ligne type de  $\mathbf{X}_{-1}$ . La GNR qui permet de calculer des estimations en une étape sera

$$\mathbf{y} - \rho\mathbf{y}_{-1} - (\mathbf{X}\dot{\boldsymbol{\beta}} - \rho\mathbf{X}_{-1}\dot{\boldsymbol{\beta}}) = (\mathbf{X} - \rho\mathbf{X}_{-1})\mathbf{b} + r\dot{\mathbf{u}}_{-1} + \text{résidus}, \quad (10.19)$$

où  $\dot{\mathbf{u}}_{-1} \equiv \mathbf{y}_{-1} - \mathbf{X}_{-1}\dot{\boldsymbol{\beta}}$ . Cette GNR se calcule directement dès que l'on connaît  $\rho$  et  $\dot{\boldsymbol{\beta}}$ , les estimations initiales convergentes de  $\rho$  et de  $\boldsymbol{\beta}$ . Si  $\mathbf{X}$  ne contient pas de variable dépendante retardée, cela est très facile. L'estimation OLS  $\tilde{\boldsymbol{\beta}}$  obtenue en régressant  $\mathbf{y}$  sur  $\mathbf{X}$  sera convergente, et une estimation convergente de  $\rho$  peut alors être obtenue en régressant  $\tilde{u}_t$  sur  $\tilde{u}_{t-1}$  pour  $t = 2$  à  $n$ . Mais si par contre  $\mathbf{X}$  contient au moins une variable dépendante retardée, cette approche simple n'opérera pas, parce que l'estimation OLS  $\tilde{\boldsymbol{\beta}}$  ne sera pas convergente. Nous traitons maintenant ce problème.

Le modèle de régression non linéaire (10.18) est un cas particulier du modèle de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{y}_{-1} + \mathbf{X}_{-1}\boldsymbol{\gamma} + \varepsilon, \quad (10.20)$$

où l'on impose la contrainte  $\boldsymbol{\gamma} = -\rho\boldsymbol{\beta}$ . Plus tard, dans la Section 10.9, nous servirons de cela pour tester l'adéquation d'une spécification AR(1). Pour l'instant, comme Durbin (1960), nous l'utiliserons simplement pour obtenir une estimation convergente de  $\rho$ . Il semblerait que l'on puisse obtenir des estimations convergentes de  $\boldsymbol{\beta}$  et de  $\rho$  en estimant (10.20) par OLS. Cela sera hélas rarement le cas, parce que nombreux seront les coefficients dans (10.20) qui ne seront pas identifiables. Par exemple, si  $\mathbf{X}$  contient un terme constant, l'un des éléments de  $\boldsymbol{\beta}$  sera le coefficient associé à la constante et l'un des éléments de  $\boldsymbol{\gamma}$  sera le coefficient associé à cette constante retardée; à l'évidence, ces deux paramètres ne sont pas identifiables séparément.

Il est aisé d'obtenir une estimation convergente de  $\rho$  à partir de (10.20) si tous les retards de la variable dépendante compris dans  $\mathbf{X}$  sont supérieurs à 1; le coefficient estimé associé à  $\mathbf{y}_{-1}$  y pourvoit. Hélas le cas où  $\mathbf{y}_{-1}$  est compris dans  $\mathbf{X}$  est vraisemblablement le plus commun. Pour expliquer les difficultés qui surviennent dans ce cas, nous allons supposer que le modèle initial est

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t. \quad (10.21)$$

Alors le modèle que nous désirons estimer en réalité, (10.18), est

$$y_t = \beta_0(1 - \rho) + (\rho + \beta_1)y_{t-1} - \rho\beta_1 y_{t-2} + \beta_2 z_t - \rho\beta_2 z_{t-1} + \varepsilon_t, \quad (10.22)$$

et l'on peut écrire le modèle non contraint (10.22) comme

$$y_t = \delta_0 + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \delta_3 z_t + \delta_4 z_{t-1} + \varepsilon_t. \quad (10.23)$$

Remarquons que le modèle non contraint a cinq coefficients de régression, qu'il faut comparer aux quatre coefficients du modèle contraint (10.22). On obtient  $\hat{\delta}_0$ ,  $\hat{\delta}_1$ ,  $\hat{\delta}_2$ ,  $\hat{\delta}_3$  et  $\hat{\delta}_4$  en estimant (10.23) par OLS, et il s'agit d'estimations convergentes des paramètres  $\delta_0$ ,  $\delta_1$ ,  $\delta_2$ ,  $\delta_3$  et  $\delta_4$ . Ces derniers sont en relation avec  $\rho$  et les  $\beta_i$  à travers les équations

$$\delta_0 = \beta_0(1 - \rho); \quad \delta_1 = \rho + \beta_1; \quad \delta_2 = -\rho\beta_1; \quad \delta_3 = \beta_2; \quad \delta_4 = -\rho\beta_2. \quad (10.24)$$

Il existe de nombreuses manières d'obtenir une estimation convergente de  $\rho$  en utilisant ces équations. La plus simple consiste à remplacer  $\beta_2$  dans la dernière équation de (10.24) par sa valeur dans l'avant dernière, pour obtenir le résultat

$$\hat{\rho} = -\hat{\delta}_4 / \hat{\delta}_3. \quad (10.25)$$

A condition que  $|\hat{\rho}| < 1$ , cette estimation convergente de  $\rho$  peut s'utiliser pour obtenir une estimation convergente de  $\beta$  en calculant  $\mathbf{y}^*(\hat{\rho})$  et  $\mathbf{X}^*(\hat{\rho})$  et en régressant le vecteur sur la matrice pour obtenir  $\hat{\beta}$ . On peut ensuite calculer des estimations en une étape en utilisant la régression de Gauss-Newton (10.19). Evidemment, puisque dans de nombreux cas le modèle originel aura plus d'un régresseur semblable à  $z_t$ , il existera de nombreux moyens d'obtenir des estimations convergentes de  $\rho$ . Cela introduit un élément d'arbitraire dans la procédure d'estimation en une étape, ce qui explique que de telles procédures soient peu employées.

Une approche alternative consiste à estimer le modèle initial par variables instrumentales de façon à obtenir une estimation convergente de  $\beta$ , estimation pour laquelle les instruments seraient les variables dépendantes retardées. On peut ensuite utiliser ces estimations pour produire une estimation convergente de  $\rho$  en régressant les résidus sur les résidus retardés de façon habituelle et, par la suite, estimer (10.19) pour obtenir des estimations en une étape qui seront asymptotiquement équivalentes aux estimations NLS. C'est l'approche

adoptée par Hatanaka (1974), qui simplifie aussi légèrement (10.19) en ne soustrayant pas  $\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\rho}\mathbf{X}_{-1}\hat{\boldsymbol{\beta}}$  de la régressande. La conséquence de cette simplification est que les estimations en une étape de  $\boldsymbol{\beta}$  sont désormais  $\hat{\mathbf{b}}$  plutôt que  $\hat{\mathbf{b}} + \hat{\boldsymbol{\beta}}$ . Tout comme les autres procédures dont nous avons discuté, la procédure de Hatanaka implique un élément arbitraire de décision, parce que les estimations convergentes initiales dépendront des instruments utilisés.

## 10.4 ECARTS TYPES ET MATRICES DE COVARIANCE

Il peut falloir beaucoup de soins pour obtenir des estimations valables de la matrice de covariance des estimations des paramètres dans le cas d'un modèle avec autocorrélation. Si l'on utilise les moindres carrés non linéaires pour estimer le modèle de régression non linéaire (10.12) qui provient de la transformation du modèle initial  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}$  en un modèle à aléas AR(1), la façon évidente d'estimer la matrice de covariance de  $\hat{\rho}$  et  $\hat{\boldsymbol{\beta}}$  est d'employer la régression de Gauss-Newton. Dans ce cas, la GNR sera

$$\mathbf{y} - \hat{\mathbf{x}} - \hat{\rho}(\mathbf{y}_{-1} - \hat{\mathbf{x}}_{-1}) = (\hat{\mathbf{X}} - \hat{\rho}\hat{\mathbf{X}}_{-1})\mathbf{b} + r(\mathbf{y}_{-1} - \hat{\mathbf{x}}_{-1}) + \text{résidus} \quad (10.26)$$

et elle est valable pour les observations allant de 2 à  $n$ . La régressande correspond au vecteur des résidus de l'estimation de (10.12) par moindres carrés non linéaires. Il y a  $k+1$  régresseurs, un associé à chacun des  $k$  éléments de  $\boldsymbol{\beta}$  et un correspondant à  $\rho$ . Dans le cas linéaire, le régresseur correspondant à  $\beta_i$  est simplement la  $i^{\text{ième}}$  colonne de  $\mathbf{X}^*(\hat{\rho})$ , et le régresseur correspondant à  $\rho$  est le vecteur  $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$  retardé une fois. Remarquons que ce dernier régresseur *n'est pas* simplement le vecteur retardé des résidus de l'estimation OLS du modèle initial sans autocorrélation, parce que  $\hat{\boldsymbol{\beta}}$  est le vecteur des paramètres estimés du modèle (10.12) corrigé pour tenir compte des aléas AR(1).

Les résultats des Chapitres 5 et 6 montrent clairement que la GNR (10.26) produira une estimation de la matrice de covariance asymptotiquement valable, et c'est celle que calcule la plupart des progiciels de moindres carrés non linéaires. Par conséquent, si le modèle de départ (10.11) est non linéaire, ou si l'on utilise les NLS alors que le modèle est linéaire, la construction d'une estimation valable de la matrice de covariance ne pose aucun problème. Cependant, l'estimation générée par (10.26) *n'est pas* l'estimation que des exécutions des procédures de Cochrane-Orcutt et Hildreth-Lu produiraient. Comme elles sont souvent mises en œuvre, ces procédures rapportent systématiquement une matrice de covariance estimée pour  $\hat{\boldsymbol{\beta}}$  à partir de la régression finale de  $\mathbf{y}^*(\hat{\rho})$  sur  $\mathbf{X}^*(\hat{\rho})$ , c'est-à-dire la régression

$$\mathbf{y} - \hat{\rho}\mathbf{y}_{-1} = (\mathbf{X} - \hat{\rho}\mathbf{X}_{-1})\boldsymbol{\beta} + \text{résidus}, \quad (10.27)$$

qui engendre le vecteur de coefficients NLS  $\hat{\boldsymbol{\beta}}$ . La matrice de covariance estimée qui en résulte, calculée à partir des  $n-1$  observations, sera

$$\frac{SSR(\hat{\boldsymbol{\beta}}, \hat{\rho})}{n-k-1} (\mathbf{X}^{*\top}(\hat{\rho})\mathbf{X}^*(\hat{\rho}))^{-1}, \quad (10.28)$$

qui ne sera valable, en général, que *conditionnellement* à  $\hat{\rho}$ . Puisque  $\rho$  a été effectivement estimé, nous voulons une estimation de la matrice de covariance valable de façon non conditionnelle. Comme nous le démontrons maintenant, (10.28) peut ou pas fournir une telle estimation de la matrice de covariance.

Dans le cas du modèle linéaire, la GNR (10.26) sera

$$\mathbf{y} - \hat{\rho}\mathbf{y}_{-1} - (\mathbf{X} - \hat{\rho}\mathbf{X}_{-1})\hat{\boldsymbol{\beta}} = (\mathbf{X} - \hat{\rho}\mathbf{X}_{-1})\mathbf{b} + r\hat{\mathbf{u}}_{-1} + \text{résidus}, \quad (10.29)$$

où  $\hat{\mathbf{u}}_{-1} \equiv \mathbf{y}_{-1} - \mathbf{X}_{-1}\hat{\boldsymbol{\beta}}$ . Remarquons que (10.27) et (10.29) ont exactement les mêmes résidus. Dans (10.29), la régressande est orthogonale à tous les régresseurs, et par conséquent le vecteur des résidus est tout simplement la régressande. Les résidus de (10.27) ont une forme algébrique comparable à la régressande de (10.29) et les deux vecteurs coïncident parce que la valeur de  $\boldsymbol{\beta}$  que l'on utilise dans (10.27) est précisément  $\hat{\boldsymbol{\beta}}$ .

La matrice de covariance estimée pour  $\hat{\boldsymbol{\beta}}$  à partir de la GNR (10.29) correspondra au bloc supérieur gauche de dimension  $k \times k$  de la matrice

$$\frac{SSR(\hat{\boldsymbol{\beta}}, \hat{\rho})}{n - k - 2} \begin{bmatrix} \mathbf{X}^{*\top}(\hat{\rho})\mathbf{X}^*(\hat{\rho}) & \mathbf{X}^{*\top}(\hat{\rho})\hat{\mathbf{u}}_{-1} \\ \hat{\mathbf{u}}_{-1}^\top \mathbf{X}^*(\hat{\rho}) & \hat{\mathbf{u}}_{-1}^\top \hat{\mathbf{u}}_{-1} \end{bmatrix}^{-1}. \quad (10.30)$$

Les premiers facteurs multiplicatifs dans (10.28) et (10.30) diffèrent seulement du nombre de degrés de liberté au dénominateur de l'estimation de  $\omega^2$ ; les numérateurs sont identiques parce qu'à la fois (10.27) et (10.29) ont la même somme des résidus au carré.<sup>2</sup> C'est la différence entre les seconds facteurs multiplicatifs qui importe. Le second facteur dans (10.28) est l'inverse de la matrice  $\mathbf{X}^{*\top}(\hat{\rho})\mathbf{X}^*(\hat{\rho})$ , alors que le second facteur dans l'estimation de la matrice de covariance à partir de la GNR est le bloc supérieur gauche de dimension  $k \times k$  de l'inverse d'une matrice de dimension  $(k + 1) \times (k + 1)$ . A condition que  $\mathbf{X}$  ne contienne pas de variable dépendante retardée,  $\mathbf{u}_{-1}$  sera indépendant de  $\mathbf{X}$ , ce qui implique que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{\mathbf{X}^{*\top}(\hat{\rho})\hat{\mathbf{u}}_{-1}}{n - 1} \right) = \mathbf{0}.$$

Ainsi  $(n - 1)^{-1}$  fois la matrice dans (10.30) sera asymptotiquement bloc-diagonale, et son inverse sera par conséquent asymptotiquement égale à

$$\begin{aligned} & \begin{bmatrix} (n - 1)^{-1} \mathbf{X}^{*\top}(\hat{\rho})\mathbf{X}^*(\hat{\rho}) & \mathbf{0} \\ \mathbf{0}^\top & (n - 1)^{-1} \hat{\mathbf{u}}_{-1}^\top \hat{\mathbf{u}}_{-1} \end{bmatrix}^{-1} \\ &= (n - 1) \begin{bmatrix} (\mathbf{X}^{*\top}(\hat{\rho})\mathbf{X}^*(\hat{\rho}))^{-1} & \mathbf{0} \\ \mathbf{0}^\top & (\hat{\mathbf{u}}_{-1}^\top \hat{\mathbf{u}}_{-1})^{-1} \end{bmatrix}. \end{aligned} \quad (10.31)$$

<sup>2</sup> Souvenons-nous que  $\omega^2$  est la variance des aléas  $\varepsilon_t$  qui apparaissent dans la régression non linéaire (10.12).



Cela montre clairement que (10.28) fournira une estimation valable du bloc supérieur gauche de dimension  $k \times k$  de l'inverse de (10.30).

A condition que  $\mathbf{X}$  ne contienne aucune variable dépendante retardée (ni aucune autre variable pouvant être corrélée à  $\mathbf{u}_{-1}$ ), la régression (10.27) produira une estimation asymptotiquement valable de la matrice de covariance de  $\hat{\beta}$ . D'autre part, si  $\mathbf{X}$  *contient* des variables dépendantes retardées, ou si elle n'est pas indépendante des aléas retardés  $\mathbf{u}_{-1}$  pour une toute autre raison, l'estimation conditionnelle (10.28) de la matrice de covariance *ne sera pas* valable. Avec de nombreux progiciels de régression, la matrice de covariance des procédures de Cochrane-Orcutt et de Hildreth-Lu qui sera rapportée sera donc non valable dans de nombreux cas. On peut soit utiliser la GNR (10.29) par ses propres moyens, soit employer les moindres carrés non linéaires dès le départ afin que le progiciel de régression le fasse.

Lorsque l'estimation de la matrice de covariance conditionnelle est valable, les écarts types fournis sont toujours trop faibles (asymptotiquement). En réalité, l'estimation de la matrice de covariance produite par la GNR (10.29) pour les estimations de  $\beta$  diffère de celle produite par (10.27) d'une matrice définie positive, en ignorant que les degrés de liberté sont différents. Pour s'en rendre compte, notons que la GNR (10.29) a les mêmes régresseurs que (10.27), plus un régresseur supplémentaire,  $\hat{\mathbf{u}}_{-1}$ . Si l'on applique le Théorème FWL à (10.29), nous voyons que l'estimation de la matrice de covariance qui en découle est identique à celle découlant d'une régression pour laquelle toutes les variables sont projetées sur le complément orthogonal de  $\hat{\mathbf{u}}_{-1}$ . Les résidus ne sont pas modifiés par la projection orthogonale et sont donc identiques à ceux de (10.27), ainsi que nous l'avons vu plus haut. La différence entre les estimateurs de la matrice de covariance de  $\hat{\beta}$  à partir de (10.29) et de (10.27) est donc proportionnelle à

$$(\mathbf{X}^{*\top}(\hat{\rho})\mathbf{M}_{\hat{\mathbf{u}}_{-1}}\mathbf{X}^*(\hat{\rho}))^{-1} - (\mathbf{X}^{*\top}(\hat{\rho})\mathbf{X}^*(\hat{\rho}))^{-1}, \quad (10.32)$$

excepté pour un effet asymptotiquement négligeable dû à une différence de degrés de liberté entre les facteurs. Si l'on soustrait les inverses des deux matrices de (10.32) dans l'ordre opposé, on obtient la matrice

$$\mathbf{X}^{*\top}(\hat{\rho})\mathbf{P}_{\hat{\mathbf{u}}_{-1}}\mathbf{X}^*(\hat{\rho}),$$

qui est à l'évidence semi-définie positive. Il suit d'après un résultat démontré dans l'Annexe A que (10.32) est elle-même une quantité semi-définie positive. Si  $\hat{\mathbf{u}}_{-1}$  est substantiellement corrélé avec les colonnes de  $\mathbf{X}^*(\hat{\rho})$ , l'estimation erronée de la variance obtenue à partir de (10.27) peut être beaucoup plus faible que l'estimation correcte obtenue à partir de la GNR (10.29).

Les régressions de Gauss-Newton (10.26) et (10.29) produisent des écarts types estimés aussi bien pour  $\hat{\rho}$  que pour  $\hat{\beta}$ . Si la matrice de covariance est asymptotiquement bloc-diagonale entre  $\rho$  et  $\beta$ , on voit à partir de (10.31) que

la variance asymptotique de  $n^{1/2}(\hat{\rho} - \rho_0)$  sera égale à

$$\omega^2 \operatorname{plim}_{n \rightarrow \infty} \left( \frac{\hat{\mathbf{u}}_{-1}^\top \hat{\mathbf{u}}_{-1}}{n-1} \right)^{-1} = \omega^2 \left( \frac{1 - \rho_0^2}{\omega^2} \right) = 1 - \rho_0^2. \quad (10.33)$$

Ainsi, dans ce cas spécial, la variance de  $\hat{\rho}$  peut être estimée par

$$\frac{1 - \hat{\rho}^2}{n-1}. \quad (10.34)$$

Il peut paraître curieux que ni la variance asymptotique  $1 - \rho_0^2$  ni l'estimation (10.34) ne dépendent de  $\omega^2$ . Après tout, nous nous attendons normalement à ce que la variance de l'estimateur d'un paramètre d'une fonction de régression soit proportionnelle à la variance des aléas. La raison pour laquelle la variance de  $\hat{\rho}$  ne dépend pas de  $\omega^2$  est que  $\omega^2$  l'influence en réalité de deux façons différentes, et ces deux influences se compensent mutuellement. On peut apercevoir cette propriété dans l'expression centrale de (10.33). La variance de  $u_{t-1}$  est directement proportionnelle à  $\omega^2$ . Ainsi, au fur et à mesure que  $\omega^2$  croît, le *rapport* de la variabilité de  $\varepsilon_t$  à celle du régresseur  $\hat{u}_{t-1}$  dans la GNR (10.29) reste constant. Etant donné que c'est précisément ce rapport qui importe pour la variance de l'estimation du coefficient, cette dernière ne dépend absolument pas de  $\omega^2$ .

L'usage de la GNR pour le calcul de la matrice de covariance de  $(\hat{\beta}, \hat{\rho})$  dans une régression linéaire avec erreurs AR(1) a été préconisé par Davidson et MacKinnon (1980). Une approche fondamentalement différente, qui était à la fois plus difficile et moins générale, fut suggérée bien avant par Cooper (1972a). Les avantages d'une analyse fondée sur la régression de Gauss-Newton sont évidents si l'on compare l'approche adoptée par Cooper avec le traitement que nous venons de donner.

## 10.5 PROCESSUS AR D'ORDRE SUPÉRIEUR

Bien que le processus AR(1) (10.01) soit de loin le plus populaire dans les études économétriques appliquées, il existe de nombreux autres processus aléatoires qui pourraient être raisonnablement utilisés pour décrire l'évolution des aléas à travers le temps. Tout ce qui ressemblerait à un traitement complet de ce thème nous conduirait très loin, dans la vaste littérature des **méthodes des séries temporelles**. Cette littérature, qui évolua de façon assez indépendante par rapport à l'économétrie et qui a influencé substantiellement cette discipline ces dernières années, traite des nombreux aspects de la modélisation des séries temporelles mais tout particulièrement lorsque les variables des modèles ne dépendent *que* de leurs propres valeurs passées (ou du moins au début). De tels modèles sont à l'évidence adéquats pour décrire l'évolution de nombreux systèmes physiques et peuvent également se révéler

adéquats pour des systèmes économiques. Toutefois, la plupart des usages des méthodes des séries temporelles en économétrie concerne la modélisation de l'évolution des aléas associés à des modèles de régression plus classiques, et nous ne traiterons ici que cet aspect des méthodes de séries temporelles. Une référence classique sur les techniques de séries temporelles est Box et Jenkins (1976), certains ouvrages qui pourraient être plus accessibles aux économistes sont ceux de Harvey (1981, 1989) et Granger et Newbold (1986). Enfin, Granger et Watson (1984) ont réalisé une étude sur les méthodes des séries temporelles pour les économètres.

Le processus AR(1) (10.01) est en vérité un cas particulier du processus **autorégressif d'ordre  $p$** , ou **AR( $p$ )**

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.35)$$

dans lequel  $u_t$  dépend autant de ses  $p$  valeurs retardées que de  $\varepsilon_t$ . On peut écrire le processus AR( $p$ ) de façon plus ramassée comme

$$(1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_p L^p) u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.36)$$

où  $L$  désigne l'**opérateur retard**. L'opérateur retard  $L$  possède la propriété de retarder d'une période l'indice temporel d'une quantité lorsque celle-ci est multipliée par  $L$ . Ainsi

$$L u_t = u_{t-1}, \quad L^2 u_t = u_{t-2}, \quad L^p u_t = u_{t-p},$$

et ainsi de suite. L'expression entre parenthèses dans (10.36) est un polynôme en  $L$ , avec les coefficients 1 et  $-\rho_1, \dots, -\rho_p$ . Si l'on définit  $A(L, \boldsymbol{\rho})$  comme étant égal à ce polynôme, où  $\boldsymbol{\rho}$  représente le vecteur  $[\rho_1 \vdots \rho_2 \vdots \cdots \vdots \rho_p]$ , on peut écrire (10.36) sous la forme compacte

$$A(L, \boldsymbol{\rho}) u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.37)$$

Pour les mêmes raisons qui nous poussent à imposer la condition  $|\rho_1| < 1$  sur les processus AR(1) de façon à assurer leur stationnarité, nous voudrions imposer des conditions de stationnarité sur les processus AR( $p$ ) plus généraux. On peut exprimer la condition de stationnarité pour de tels processus de plusieurs manières différentes; l'une d'elles consiste à dire que toutes les racines de l'équation polynomiale en  $z$ ,

$$A(z, \boldsymbol{\rho}) \equiv 1 - \rho_1 z - \rho_2 z^2 - \cdots - \rho_p z^p = 0 \quad (10.38)$$

doivent se situer **à l'extérieur du cercle de rayon 1**, ce qui signifie simplement que toutes les racines de (10.38) doivent être supérieures à 1 en valeur absolue. Cette condition peut mener à des contraintes sur  $\boldsymbol{\rho}$  assez délicates pour les processus AR( $p$ ).

Il est rarement utile de spécifier un processus  $AR(p)$  d'ordre élevé (c'est-à-dire un processus avec un  $p$  élevé) lorsque l'on essaie de modéliser les aléas associés à un modèle de régression. Le processus  $AR(2)$  est de loin plus souple, mais aussi plus compliqué, que le processus  $AR(1)$ ; c'est souvent tout ce dont on a besoin lorsque ce dernier est trop contraignant. On voit clairement la complexité supplémentaire du processus  $AR(2)$ . Par exemple, la variance de  $u_t$ , sous l'hypothèse de stationnarité, est

$$\sigma^2 = \frac{\omega^2(1 - \rho_2)}{(1 + \rho_2)^3 - (1 + \rho_2)\rho_1^2},$$

qui est sensiblement plus compliquée que l'expression (10.02) correspondant au cas  $AR(1)$ , et la stationnarité implique désormais que trois conditions soient vérifiées:

$$\rho_1 + \rho_2 < 1; \quad \rho_2 - \rho_1 < 1; \quad \rho_2 > -1. \quad (10.39)$$

Les conditions (10.39) définissent un **triangle de stationnarité**. Les sommets de ce triangle ont pour coordonnées  $(-2, -1)$ ,  $(2, -1)$ , et  $(0, 1)$ . A condition que le point  $(\rho_1, \rho_2)$  n'appartienne pas au triangle, le processus  $AR(2)$  sera stationnaire.

Les processus autorégressifs d'ordre supérieur à 2 apparaissent assez fréquemment avec des données temporelles qui présentent des variations saisonnières. Il n'est pas rare, par exemple, que les aléas de modèles estimés à l'aide de données trimestrielles suivent apparemment le **processus  $AR(4)$  simple**

$$u_t = \rho_4 u_{t-4} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.40)$$

dans lequel l'aléa de la période  $t$  dépend de l'aléa du même trimestre de l'année précédente, mais pas des aléas intermédiaires. Autre possibilité, les aléas peuvent sembler suivre un processus  $AR$  combinant premier et quatrième ordres

$$(1 - \rho_1 L)(1 - \rho_4 L^4)u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2) \quad (10.41)$$

ou, en développant le polynôme de gauche,

$$(1 - \rho_1 L - \rho_4 L^4 + \rho_1 \rho_4 L^5)u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2).$$

C'est un cas contraint particulier d'un processus  $AR(5)$ , mais avec seulement deux paramètres au lieu de cinq. Nous discuterons au cours du Chapitre 19 des différentes façons de modéliser la saisonnalité, et parmi elles les **processus  $AR$  saisonniers** tels que (10.40) et (10.41).

Il est clair que l'estimation d'un modèle de régression dont les aléas suivent un processus  $AR(p)$  n'est pas fondamentalement différente de celle du même processus où les aléas suivraient un processus  $AR(1)$ . Ainsi, par exemple, si l'on veut estimer le modèle

$$y_t = x_t(\beta) + u_t, \quad (1 - \rho_1 L)(1 - \rho_4 L^4)u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2),$$

il suffit simplement de le transformer en

$$y_t = x_t(\beta) + \rho_1(y_{t-1} - x_{t-1}(\beta)) + \rho_4(y_{t-4} - x_{t-4}(\beta)) \\ - \rho_1\rho_4(y_{t-5} - x_{t-5}(\beta)) + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2),$$

d'éliminer les *cinq* premières observations, et d'utiliser les moindres carrés non linéaires. Comme dans le cas AR(1), la matrice de covariance de  $(\hat{\rho}, \hat{\beta})$  peut alors être estimée à l'aide de la régression de Gauss-Newton. La mise à l'écart de cinq observations peut nous rendre mal à l'aise, en particulier si la taille de l'échantillon est faible, mais elle est certainement valable asymptotiquement, et puisque tous nos résultats sur les moindres carrés sont asymptotiques, il n'y a pas de raison majeure qui nous empêche de le faire. Nous discuterons des approches alternatives dans la prochaine section.

## 10.6 OBSERVATIONS INITIALES DES MODÈLES À ALÉAS AR

Jusqu'à présent, lorsque nous avons transformé les modèles de régression à aléas autorégressifs de façon à obtenir des aléas bruits blancs, il nous a juste fallu mettre à l'écart autant d'observations du début de l'échantillon que nécessaire pour faire du modèle transformé un modèle de régression non linéaire. Bien que cela soit à l'évidence valable asymptotiquement et sûrement le moyen le plus simple de procéder, les chercheurs pourraient hésiter à rejeter les informations portées par les observations initiales. Cette hésitation peut reposer sur des raisons sérieuses. Comme nous le verrons, ces observations initiales peuvent parfois porter beaucoup plus d'information sur les valeurs des paramètres que leur nombre relatif ne le laisserait supposer. Par conséquent, leur mise à l'écart peut provoquer une sérieuse perte d'efficacité.

Nous débutons par le modèle de régression non linéaire avec aléas AR(1). Comme nous l'avons vu, le processus AR(1) est le processus AR le plus simple à analyser et le plus fréquent dans les travaux empiriques. De plus, cela nous permet d'introduire tous les résultats conceptuels intéressants associés au traitement des observations initiales. Le modèle est

$$y_t = x_t(\beta) + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad |\rho| < 1. \quad (10.42)$$

L'expression (10.05) donne la matrice de covariance des  $u_t$ . On peut vérifier par une multiplication que l'inverse de cette matrice est

$$\Omega^{-1} = \frac{1}{\omega^2} \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \equiv \frac{\Delta^{-1}(\rho)}{\omega^2}. \quad (10.43)$$

De manière similaire, on peut vérifier que pour la matrice  $\boldsymbol{\eta}(\rho)$ , qui doit satisfaire la propriété

$$\boldsymbol{\eta}^\top(\rho)\boldsymbol{\eta}(\rho) \equiv \boldsymbol{\Delta}^{-1}(\rho),$$

on peut faire usage de

$$\boldsymbol{\eta}(\rho) = \begin{bmatrix} (1 - \rho^2)^{1/2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}. \quad (10.44)$$

Cette matrice transformée fut dérivée par Prais et Winsten (1954), et on y fait souvent référence en tant que **transformation de Prais-Winsten**. Ainsi, si  $\rho$  était connu, on pourrait obtenir des estimations GNLS en exécutant la régression non linéaire

$$\boldsymbol{\eta}(\rho)\mathbf{y} = \boldsymbol{\eta}(\rho)\mathbf{x}(\boldsymbol{\beta}) + \boldsymbol{\eta}(\rho)\mathbf{u}$$

ou, en adoptant un changement de notation évident,

$$\mathbf{y}^*(\rho) = \mathbf{x}^*(\boldsymbol{\beta}, \rho) + \mathbf{u}^*. \quad (10.45)$$

En postmultipliant  $\boldsymbol{\eta}(\rho)$  par  $\mathbf{y}$  et  $\mathbf{x}(\boldsymbol{\beta})$ , nous voyons que la régressande  $\mathbf{y}^*(\rho)$  et le vecteur de fonctions de régression  $\mathbf{x}^*(\rho)$  sont, respectivement

$$\begin{aligned} y_1^*(\rho) &= (1 - \rho^2)^{1/2} y_1; \\ y_t^*(\rho) &= y_t - \rho y_{t-1} \text{ pour tout } t \geq 2; \\ x_1^*(\boldsymbol{\beta}, \rho) &= (1 - \rho^2)^{1/2} x_1(\boldsymbol{\beta}); \\ x_t^*(\boldsymbol{\beta}, \rho) &= x_t(\boldsymbol{\beta}) - \rho x_{t-1}(\boldsymbol{\beta}) \text{ pour tout } t \geq 2. \end{aligned} \quad (10.46)$$

Ainsi pour les observations allant de 2 à  $n$ , le modèle transformé (10.45) est identique au modèle de régression linéaire (10.12) pour lequel on a mis à l'écart la première observation. Ce qui est différent est que (10.45) s'applique désormais à  $n$  observations au lieu de  $n - 1$ .

La régression (10.45) montre que l'on peut calculer des estimations de  $\boldsymbol{\beta}$  par GNLS, et, par extension, par GNLS faisables, en utilisant les  $n$  observations. Les GNLS faisables seront disponibles chaque fois qu'il sera possible d'obtenir une estimation convergente de  $\rho$ . Comme nous l'avons vu dans la Section 10.3, cela sera aisé si  $x_t(\boldsymbol{\beta})$  ne dépend pas des valeurs retardées de  $y_t$ , puisqu'il nous faut simplement estimer le modèle par NLS et régresser les résidus sur eux-mêmes retardés une fois. Mais si  $x_t(\boldsymbol{\beta})$  dépend des valeurs retardées de  $y_t$ , la régression (10.45) n'est pas un moyen adéquat de gérer

la première observation. On peut écrire la première observation de (10.45) comme

$$(1 - \rho^2)^{1/2} y_1 = (1 - \rho^2)^{1/2} x_1(\beta) + (1 - \rho^2)^{1/2} u_1, \quad (10.47)$$

qui montre que l'unique effet de la transformation (10.44) est de tout multiplier par  $(1 - \rho^2)^{1/2}$ . Nous avons vu dans la Section 10.2 qu'une fonction de régression qui dépend des retards de la variable dépendante est corrélée avec les aléas correspondants si ces aléas sont autocorrélés. Si c'est le cas pour  $x_1(\beta)$  et  $u_1$ , nous voyons à partir de (10.47) que cela doit également être le cas pour  $x_1^*(\beta, \rho)$  et  $u_1^*(\rho)$ . Evidemment, puisque la corrélation entre  $x_1^*(\beta, \rho)$  et  $u_1^*(\rho)$  n'affecte qu'une seule observation, il est parfaitement valable asymptotiquement de traiter la première observation de cette façon, autant qu'il est parfaitement valable de la mettre à l'écart. Il est possible, mais nullement aisé, de tenir compte convenablement de la première observation dans un modèle de régression linéaire avec une seule variable dépendante retardée et des aléas AR(1); voir Pesaran (1981). Cependant, il est beaucoup plus fréquent de simplement éliminer la première observation dans ce cas. Dans le reste de cette section, nous supposons donc que  $x_t(\beta)$  ne dépend pas des valeurs retardées de  $y_t$ .

Nous avons vu désormais la manière d'obtenir des estimations par GNLS et par GNLS faisables du modèle (10.42) qui utilisent toutes deux les  $n$  observations. Si l'on suppose de plus que les  $\varepsilon_t$  sont normalement distribués, on peut obtenir des estimations ML. Parce que celles-ci sont asymptotiquement équivalentes aux estimations GNLS, elles seront convergentes même si l'hypothèse de normalité est fausse. Les techniques qui estiment ce modèle par maximum de vraisemblance en considérant toutes les observations sont souvent appelées **estimation ML totale** ou **estimation ML exacte**.

Il existe de nombreux moyens de dériver la fonction de logvraisemblance. Le plus simple est sûrement celui-ci. Pour les observations allant de 2 à  $n$ , nous avons vu que

$$y_t = \rho y_{t-1} + x_t(\beta) - \rho x_{t-1}(\beta) + \varepsilon_t.$$

On peut combiner les éléments de cette expression de façon à exprimer  $\varepsilon_t$  en fonction de  $y_t$ . La densité de  $\varepsilon_t$  est

$$\frac{1}{\sqrt{2\pi}} \frac{1}{\omega} \exp\left(-\frac{\varepsilon_t^2}{2\omega^2}\right).$$

Puisque  $\partial \varepsilon_t / \partial y_t = 1$ , le Jacobien est égal à l'unité, et par conséquent la fonction de logvraisemblance partielle des observations allant de 2 à  $n$  est

$$\begin{aligned} \ell^{2,n}(\mathbf{y}, \beta, \rho, \omega) = & -\frac{n-1}{2} \log(2\pi) - (n-1) \log(\omega) \\ & - \frac{1}{2\omega^2} \sum_{t=2}^n (y_t - \rho y_{t-1} - x_t(\beta) + \rho x_{t-1}(\beta))^2. \end{aligned} \quad (10.48)$$

La densité de  $u_1 = y_1 - x_1(\beta)$  est

$$\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{u_1^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}} \frac{1}{\omega} (1 - \rho^2)^{1/2} \exp\left(-\frac{1 - \rho^2}{2\omega^2} u_1^2\right), \quad (10.49)$$

où la première expression décrit la densité en fonction de  $\sigma$ , et la seconde en fonction de  $\omega = \sigma(1 - \rho^2)^{1/2}$ . Ainsi la contribution de l'observation 1 à la logvraisemblance est

$$\begin{aligned} \ell^1(\mathbf{y}, \beta, \rho, \omega) = & -\frac{1}{2} \log(2\pi) - \log(\omega) + \frac{1}{2} \log(1 - \rho^2) \\ & - \frac{1 - \rho^2}{2\omega^2} (y_1 - x_1(\beta))^2. \end{aligned} \quad (10.50)$$

La combinaison de (10.48) et (10.50) produit la fonction de logvraisemblance totale

$$\begin{aligned} \ell^n(\mathbf{y}, \beta, \rho, \omega) = & -\frac{n}{2} \log(2\pi) - n \log(\omega) + \frac{1}{2} \log(1 - \rho^2) \\ & - \frac{1}{2\omega^2} \left( \sum_{t=2}^n (y_t - \rho y_{t-1} - x_t(\beta) + \rho x_{t-1}(\beta))^2 + (1 - \rho^2)(y_1 - x_1(\beta))^2 \right). \end{aligned} \quad (10.51)$$

En concentrant  $\ell^n(\mathbf{y}, \beta, \rho, \omega)$  par rapport à  $\omega$ , on obtient la fonction de logvraisemblance concentrée totale  $\ell^c(\mathbf{y}, \beta, \rho)$ :

$$C + \frac{1}{2} \log(1 - \rho^2) - \frac{n}{2} \log\left((\mathbf{y} - \mathbf{x}(\beta))^\top \mathbf{\Delta}^{-1}(\rho) (\mathbf{y} - \mathbf{x}(\beta))\right), \quad (10.52)$$

où la matrice  $\mathbf{\Delta}^{-1}(\rho)$  de dimension  $n \times n$  est explicitement définie par (10.43).

On peut maximiser la fonction (10.52) de plusieurs façons. En particulier, dans le cas où  $\mathbf{x}(\beta) = \mathbf{X}\beta$ , Beach et MacKinnon (1978a) ont proposé un algorithme comparable à la procédure itérative de Cochrane-Orcutt dont nous avons discuté dans la Section 10.3. Ils ont montré que la maximisation de  $\ell^c(\mathbf{y}, \beta, \rho)$  conditionnellement à  $\beta$  passe par la recherche de la valeur de la racine centrale d'une certaine équation du troisième degré en  $\rho$ . La formule qui le permet, combinée à la version linéaire de la régression (10.45), autorise le calcul d'une succession de valeurs  $\beta^{(0)}$ ,  $\rho^{(1)}$ ,  $\beta^{(1)}$ ,  $\rho^{(2)}$ , et ainsi de suite, qui doivent en fin de compte converger vers le maximum local de (10.52) pour les mêmes raisons que celles avancées dans le cas de la procédure itérative de Cochrane-Orcutt.

Le terme  $\frac{1}{2} \log(1 - \rho^2)$  qui apparaît à la fois dans (10.51) et dans (10.52) est un terme jacobien. Cela peut ne pas être clair d'après la manière dont nous avons dérivé la fonction de logvraisemblance. On peut imaginer que la première observation est

$$(1 - \rho^2)^{1/2} y_1 = (1 - \rho^2)^{1/2} x_1(\beta) + \varepsilon_1, \quad (10.53)$$



où  $\varepsilon_1$  est  $N(0, \omega^2)$ .<sup>3</sup> Ainsi on obtient la densité (10.49) en transformant  $\varepsilon_1$  en  $y_1$ , et le facteur jacobien  $(1 - \rho^2)^{1/2}$  provient de cette transformation. Le terme jacobien qui en résulte  $\frac{1}{2} \log(1 - \rho^2)$  joue un rôle extrêmement important dans l'estimation. Puisqu'il tend vers moins l'infini quand  $\rho$  tend vers  $\pm 1$ , sa présence au sein de la fonction de logvraisemblance garantit l'existence d'un maximum à l'intérieur de la **région de stationnarité**  $-1 < \rho < 1$ . Ainsi on a la *garantie* que l'estimation par le maximum de la vraisemblance totale produit une estimation de  $\rho$  pour laquelle le processus AR(1) des aléas est stationnaire. Cela n'est pas le cas des autres techniques d'estimation. Les techniques qui mettent la première observation à l'écart peuvent, et le font effectivement quelquefois, produire des estimations de  $\rho$  supérieures à 1 en valeur absolue. La procédure itérative de Cochrane-Orcutt (si elle converge vers le maximum global) est équivalente à la maximisation de la fonction de logvraisemblance (10.48) pour les  $n - 1$  observations uniquement, et rien n'empêche d'aboutir à un maximum pour une valeur de  $\rho$  n'appartenant pas à la région de stationnarité. C'est également le cas de la procédure itérative de Prais-Winsten qui est comparable à celle de Cochrane-Orcutt mais qui fait usage de la transformation (10.44) pour trouver  $\beta$  en fonction de  $\rho$ , de manière à minimiser  $(\mathbf{y} - \mathbf{x}(\beta))^\top \Delta^{-1}(\rho)(\mathbf{y} - \mathbf{x}(\beta))$ . Evidemment, puisque la transformation (10.44) n'a aucun sens lorsque  $|\rho| > 1$ , de telles estimations devraient être abandonnées.

On peut estimer la matrice de covariance du vecteur des estimations ML  $[\hat{\beta} : \hat{\rho} : \hat{\omega}]$  en calculant l'inverse de la matrice d'information et en l'évaluant en  $[\hat{\beta} : \hat{\rho} : \hat{\omega}]$ . Le résultat est

$$\hat{\mathbf{V}}(\hat{\beta}, \hat{\rho}, \hat{\omega}) = \begin{bmatrix} \hat{\omega}^2 (\hat{\mathbf{X}}^{*\top} \hat{\mathbf{X}}^*)^{-1} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{V}}(\hat{\rho}, \hat{\omega}) \end{bmatrix}, \quad (10.54)$$

où  $\hat{\mathbf{X}}^*$  désigne la matrice de dimension  $n \times k$  des dérivées partielles par rapport aux éléments de  $\beta$  du vecteur des fonctions non linéaires  $\mathbf{x}^*(\beta, \rho)$ , défini en (10.46), évaluée en  $(\hat{\beta}, \hat{\rho})$ , et où

$$\hat{\mathbf{V}}(\hat{\rho}, \hat{\omega}) = \begin{bmatrix} \frac{2\hat{\rho}^2}{(1 - \hat{\rho})^2} + \frac{n - 1}{1 - \hat{\rho}^2} & \frac{2\hat{\rho}}{\hat{\omega}(1 - \hat{\rho})^2} \\ \frac{2\hat{\rho}}{\hat{\omega}(1 - \hat{\rho})^2} & \frac{2n}{\hat{\omega}^2} \end{bmatrix}^{-1}.$$

La matrice de covariance estimée (10.54) est bloc-diagonale entre  $\beta$  et  $\rho$  et entre  $\beta$  et  $\omega$  (souvenons-nous que nous avons éliminé les variables dépendantes

<sup>3</sup> Notons, cependant, que  $\varepsilon_1$  n'est pas l'innovation de la première période mais plutôt une autre variable aléatoire, de distribution identique, qui dépend en réalité de toutes les innovations jusqu'à la période 1 comprise. En effet, comme on peut le voir à partir de (10.47),  $\varepsilon_1 = (1 - \rho^2)^{1/2} u_1$ .

retardées). Cependant, et contrairement à la situation des modèles de régression, elle n'est pas bloc-diagonale entre  $\rho$  et  $\omega$ . Les éléments en dehors de la diagonale dans le bloc  $(\rho, \omega)$  de la matrice d'information sont  $O(1)$ , alors que ceux situés sur la diagonale sont  $O(n)$ . Ainsi  $V(\hat{\beta}, \hat{\rho}, \hat{\omega})$  sera asymptotiquement bloc-diagonale entre  $\beta$ ,  $\rho$ , et  $\omega$ . Cela correspond à ce que nous espérions, puisque c'est seulement la première observation, asymptotiquement négligeable, qui empêche (10.54) d'être bloc-diagonale.

C'est un exercice excellent que de retrouver la matrice de covariance estimée (10.54). Il faut commencer par calculer les dérivées secondes de (10.51) par rapport à tous les paramètres du modèle pour construire la matrice Hessienne, puis calculer les espérances de son opposée pour obtenir la matrice d'information. Il faut alors remplacer les paramètres par leur estimation ML et inverser la matrice d'information pour aboutir à (10.54). Bien que cet exercice soit assez rapide et direct, il y a de nombreuses occasions de faire des erreurs. Par exemple, Beach et MacKinnon (1978a) n'ont pas envisagé toutes les espérances possibles, et, par la suite, aboutissent à une matrice de covariance estimée excessivement compliquée.

La discussion précédente montre clairement que la prise en compte de la première observation est sensiblement plus difficile que sa mise à l'écart. Même si l'on dispose d'un programme informatique adapté, de façon à rendre l'estimation assez directe, on est confronté à des ennuis sérieux lorsque l'on teste le modèle. Puisque le modèle transformé n'est plus du tout un modèle de régression, la régression de Gauss-Newton ne s'applique plus du tout et ne peut plus être utilisée pour les tests de spécification du modèle; voir les Sections 10.8 et 10.9. On pourrait bien évidemment estimer le modèle deux fois, la première fois en tenant compte de la première observation, de façon à obtenir les estimations les plus efficaces possibles, et la seconde fois en la mettant à l'écart, de façon à être en mesure de tester la spécification, mais cela engage forcément un travail supplémentaire. La question évidente qui survient est de savoir si les ennuis supplémentaires générés par la prise en compte de la première observation en valent la peine.

Il existe une littérature étendue sur ce sujet, dont les articles de Kadiyala (1968), Rao et Griliches (1969), Maeshiro (1976, 1979), Beach et MacKinnon (1978a), Chipman (1979), Spitzer (1979), Park et Mitchell (1980), Ansley et Newbold (1980), Poirier (1981), Magee (1987), et Thornton (1987). Dans de nombreux cas, conserver la première observation entraîne des estimations plus efficaces, mais de peu. Cependant, lorsque la taille de l'échantillon est faible et qu'il y a un ou plusieurs régresseurs de tendance temporelle, il peut être crucial de retenir la première observation. Dans de telles circonstances, les procédures d'estimation par ML ou par GLS utilisant la transformation qui met à l'écart la première observation peuvent être sensiblement moins efficaces que celles qui manipulent l'échantillon entier, et peut-être même moins efficaces que les OLS.

**Tableau 10.1** Données Initiales et Données Transformées

$C_t$	$T_t$	$C_t^*(.5)$	$T_t^*(.5)$	$C_t^*(.9)$	$T_t^*(.9)$
1.0	1.0	0.866	0.866	0.436	0.436
1.0	2.0	0.5	1.5	0.1	1.1
1.0	3.0	0.5	2.0	0.1	1.2
1.0	4.0	0.5	2.5	0.1	1.3
1.0	5.0	0.5	3.0	0.1	1.4
1.0	6.0	0.5	3.5	0.1	1.5
1.0	7.0	0.5	4.0	0.1	1.6
1.0	8.0	0.5	4.5	0.1	1.7
1.0	9.0	0.5	5.0	0.1	1.8
1.0	10.0	0.5	5.5	0.1	1.9

Pour comprendre pourquoi la première observation peut être cruciale dans certains cas, considérons l'exemple simple qui suit. Le modèle est

$$y_t = \beta_0 C_t + \beta_1 T_t + u_t,$$

où  $C_t$  est une constante et  $T_t$  une variable de tendance linéaire. Dans le Tableau 10.1, on a rapporté 10 observations sur  $C_t$  et  $T_t$  avant et après la transformation (10.46) pour  $\rho = 0.5$  et  $\rho = 0.9$ .

On constate d'après le Tableau 10.1 que les données transformées pour la première observation sont très différentes de celles des observations suivantes. Il en résulte que cette observation contribue grandement à fournir de l'information sur les paramètres. On peut en voir la raison en examinant les éléments diagonaux de la “matrice chapeau”  $\mathbf{P}_{C,T}$  qui projette orthogonalement sur  $\mathcal{S}(\mathbf{C}^*, \mathbf{T}^*)$ , pour diverses valeurs de  $\rho$ . Ces éléments diagonaux apparaissent dans le Tableau 10.2.

Comme nous l'avons vu dans la Section 1.6, les éléments diagonaux de la matrice chapeau mesurent l'*effet levier* des diverses observations, c'est-à-dire leur effet potentiel sur les estimations des paramètres. Ainsi nous voyons à partir du Tableau 10.2 que, lorsque  $\rho$  croît, la première observation devient de plus en plus influente relativement au reste de l'échantillon. En fait, avec  $\rho = 0.9$ , elle correspond à un point à très fort effet levier (souvenons-nous que les éléments diagonaux d'une matrice de projection orthogonale ne peuvent jamais être supérieurs à 1). D'autre part, lorsque l'on exclut la première observation, comme dans la dernière colonne du tableau, aucune observation n'est susceptible d'avoir le poids qu'avait précédemment la première observation. Ainsi il n'est pas étonnant de voir que les écarts types des estimations des paramètres prennent de très fortes valeurs lorsque l'on omet la première observation. Par exemple, lorsque  $\rho = 0.5$ , les écarts types de  $\hat{\beta}_0$  et  $\hat{\beta}_1$  sont multipliés respectivement par 1.765 et 1.533 si l'on met à l'écart l'observation 1. Si  $\rho = 0.9$ , ils sont multipliés par 8.039 et 4.578.

**Tableau 10.2** Éléments Diagonaux de  $\mathbf{P}_{C,T}$ 

$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0.5$ et $\rho = 0.9, t = 2, 10$
0.3455	0.6808	0.9847	
0.2485	0.1277	0.0598	0.3778
0.1758	0.0993	0.0693	0.2611
0.1273	0.0851	0.0805	0.1778
0.1030	0.0851	0.0932	0.1278
0.1030	0.0993	0.1075	0.1111
0.1273	0.1277	0.1234	0.1278
0.1758	0.1702	0.1409	0.1778
0.2485	0.2270	0.1600	0.2611
0.3455	0.2979	0.1807	0.3778

Cet exemple est bien évidemment un cas limite. Nous utiliserons rarement des échantillons de 10 observations, et rares sont les cas où nous désirons régresser quelque chose sur une constante et sur une variable de tendance (ou sur un régresseur qui ressemble à une variable de tendance; lorsque les données expriment une tendance chronologique très typée, on préférera souvent les transformer afin d'éliminer la tendance avant l'estimation du modèle). Quoi qu'il en soit, cet exemple montre clairement que la gestion de la première observation peut être décisive.

Le résultat de la gestion des observations initiales s'applique autant aux modèles de régression avec aléas d'ordre supérieur qu'aux modèles à aléas AR(1). Dans le cas général AR( $p$ ), le modèle est (sous l'hypothèse de normalité)

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t = \sum_{j=1}^p \rho_j u_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, \omega^2),$$

où les  $\rho_j$  sont supposés satisfaire la condition de stationnarité qui veut que les racines de l'équation polynômiale (10.38) se situent à l'extérieur du cercle de rayon 1. Comme dans le contexte AR(1), le moyen le plus simple de dériver la fonction de logvraisemblance pour ce modèle est de la traiter comme la somme de deux fonctions de logvraisemblance, l'une valant pour les  $p$  premières observations et l'autre valant pour les observations allant de  $p+1$  à  $n$  conditionnellement aux  $p$  premières observations. La seconde fonction est

$$\begin{aligned} \ell^{p+1,n}(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\rho}, \omega) = & -\frac{n-p}{2} \log(2\pi) - (n-p) \log(\omega) \\ & - \frac{1}{2\omega^2} \sum_{t=p+1}^n \left( y_t - x_t(\boldsymbol{\beta}) - \sum_{j=1}^p \rho_j (y_{t-j} - x_{t-j}(\boldsymbol{\beta})) \right)^2. \end{aligned} \quad (10.55)$$

Ceci est à l'évidence très comparable à (10.48) pour le cas AR(1).

La fonction de logvraisemblance des  $p$  premières observations est le logarithme de la fonction de densité jointe du vecteur  $\mathbf{y}^p$ , composé des  $p$  premières observations des  $y_t$ . Si  $\omega^2 \mathbf{\Delta}_p$  désigne la matrice de covariance de dimension  $p \times p$  des  $p$  premiers  $u_t$  et si  $\mathbf{x}^p(\boldsymbol{\beta})$  désigne les  $p$  premières observations des  $x_t(\boldsymbol{\beta})$ , cette fonction sera

$$\begin{aligned} \ell^p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\rho}, \omega) = & -\frac{p}{2} \log(2\pi) - p \log(\omega) + \frac{1}{2} \log |\mathbf{\Delta}_p^{-1}| \\ & - \frac{1}{2\omega^2} (\mathbf{y}^p - \mathbf{x}^p(\boldsymbol{\beta}))^\top \mathbf{\Delta}_p^{-1} (\mathbf{y}^p - \mathbf{x}^p(\boldsymbol{\beta})). \end{aligned} \quad (10.56)$$

Si  $p = 1$ ,  $|\mathbf{\Delta}_p^{-1}| = \mathbf{\Delta}_p^{-1} = 1 - \rho^2$ . Ainsi il faut voir (10.50) comme un cas particulier de (10.56).

La fonction de logvraisemblance totale est la somme de (10.55) et (10.56). Comme dans le cas AR(1), la présence du terme jacobien  $\frac{1}{2} \log |\mathbf{\Delta}_p^{-1}|$  garantit que la fonction possédera au moins un maximum compris dans la région de stationnarité. Cependant, il rend l'évaluation et la maximisation de la fonction beaucoup plus difficiles. Certains auteurs (dont Box et Jenkins (1976)) ont à cette occasion suggéré de l'ignorer et de maximiser le reste de la fonction. Les références que l'on peut citer en matière d'estimation des modèles à aléas AR( $p$ ) sont Ansley (1979), Kendall, Stuart, et Ord (1983), et Granger et Newbold (1986). Beach et MacKinnon (1978b) discutent du cas AR(2) en détail.

## 10.7 PROCESSUS MOYENNE MOBILE ET ARMA

Les processus autorégressifs ne sont pas les seules façons de modéliser des séries temporelles stationnaires. L'autre genre de processus stochastique de base est le processus **moyenne mobile**, ou **MA**. Le processus moyenne mobile le plus simple est le processus **moyenne mobile au premier ordre**, ou **MA(1)**

$$u_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1}, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.57)$$

pour lequel l'aléa  $u_t$  est à proprement parler une moyenne mobile de deux innovations successives  $\varepsilon_t$  et  $\varepsilon_{t-1}$ . Ainsi  $\varepsilon_t$  affecte à la fois  $u_t$  et  $u_{t+1}$  mais pas  $u_{t+j}$  pour  $j > 1$ . Le processus MA( $q$ ) plus général peut s'écrire soit comme

$$u_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \cdots + \alpha_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \text{IID}(0, \omega^2)$$

soit comme

$$u_t = (1 + \alpha_1 L + \cdots + \alpha_q L^q) \varepsilon_t \equiv B(L, \boldsymbol{\alpha}) \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.58)$$

si l'on use de la notation avec l'opérateur retard, et où  $\boldsymbol{\alpha} \equiv [\alpha_1 \vdots \alpha_2 \vdots \cdots \vdots \alpha_q]$ .

Les processus MA d'ordre fini sont nécessairement stationnaires, puisque chaque  $u_t$  est une somme pondérée d'un nombre fini d'innovations  $\varepsilon_t, \varepsilon_{t-1}, \dots$ . Ainsi nous n'avons pas à imposer la stationnarité. Cependant, il nous faut imposer une **condition d'inversibilité** si l'on veut que  $\alpha$  soit identifié par les données. Dans le cas MA(1), cette condition s'exprime par  $|\alpha_1| \leq 1$ . La raison pour laquelle nous avons besoin d'une condition d'inversibilité est qu'il y aura autrement, en général, plus d'une valeur de  $\alpha$  qui produira des  $u_t$  dont les propriétés statistiques sont identiques. Par exemple, on peut montrer que le processus MA(1) (10.57) avec  $\alpha_1 = \gamma$ ,  $-1 < \gamma < 1$ , peut être indiscernable du processus MA(1) avec  $\alpha_1 = 1/\gamma$ . Nous discuterons de cela plus tard, lorsque nous traiterons l'estimation par ML des modèles à erreurs MA(1). La condition d'inversibilité pour un processus MA( $q$ ) consiste à ce que les racines du polynôme

$$B(z, \alpha) \equiv 1 + \alpha_1 z + \alpha_2 z^2 + \dots + \alpha_q z^q = 0 \quad (10.59)$$

se situent à l'extérieur du cercle de rayon un. Cette condition sur (10.59) est formellement identique à celle sur (10.38) qui assure que le processus AR( $p$ ) est stationnaire.

Il est assez direct de calculer la matrice de covariance pour un processus moyenne mobile. Par exemple, dans le contexte MA(1) la variance de  $u_t$  est évidemment

$$\sigma^2 \equiv E(\varepsilon_t + \alpha_1 \varepsilon_{t-1})^2 = \omega^2 + \alpha_1^2 \omega^2 = (1 + \alpha_1^2) \omega^2,$$

la covariance de  $u_t$  et  $u_{t-1}$  est

$$E(\varepsilon_t + \alpha_1 \varepsilon_{t-1})(\varepsilon_{t-1} + \alpha_1 \varepsilon_{t-2}) = \alpha_1 \omega^2,$$

et celle entre  $u_t$  et  $u_{t-j}$  pour  $j > 1$  est nulle. La matrice de covariance de  $\mathbf{u}$  est

$$\omega^2 \begin{bmatrix} 1 + \alpha_1^2 & \alpha_1 & 0 & \dots & 0 & 0 & 0 \\ \alpha_1 & 1 + \alpha_1^2 & \alpha_1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \alpha_1 & 1 + \alpha_1^2 & \alpha_1 \\ 0 & 0 & 0 & \dots & 0 & \alpha_1 & 1 + \alpha_1^2 \end{bmatrix}. \quad (10.60)$$

La structure de cette matrice de covariance est très simple. Remarquons que la corrélation entre deux aléas successifs varie seulement entre  $-0.5$  et  $0.5$ , puisque ces valeurs sont les valeurs minimale et maximale de  $\alpha_1/(1 + \alpha_1^2)$ , atteintes lorsque  $\alpha_1 = -1$  et  $\alpha_1 = 1$ , respectivement. Il est donc évident d'après l'examen de (10.60) qu'un processus MA(1) n'est pas adéquat lorsque la corrélation observée entre des résidus successifs est forte en valeur absolue.

Bien que les processus moyenne mobile ne soient pas aussi largement utilisés en économétrie que les processus autorégressifs, sans doute parce que ces derniers sont moins difficiles à estimer, il y a des circonstances dans lesquelles les processus MA s'imposent naturellement. Considérons le problème de l'estimation d'une équation qui explique la valeur d'un instrument financier tel que les Bons du Trésor à échéance de 90 jours ou des contrats en commerce extérieur à terme de 3 mois. Si l'on utilisait des données mensuelles, alors toute innovation dont l'occurrence est au mois  $t$  affecterait la valeur des instruments qui arrivent à échéance aux mois  $t$ ,  $t+1$ , et  $t+2$  mais n'affecterait pas directement la valeur des instruments qui arrivent à terme plus tard parce que ces derniers n'ont pas encore été émis. Cela suggère que l'aléa devrait être modélisé par un processus MA(2); consulter Frankel (1980) et Hansen et Hodrick (1980). Les aléas moyenne mobile apparaissent également lorsque les données sont collectées à l'aide d'un sondage où certains individus identiques sont questionnés à des périodes successives, tel que les sondages sur la main-d'œuvre salariée aux USA comme au Canada, utilisés pour estimer les taux de chômage; consulter Hausman et Watson (1985).

Il est généralement beaucoup plus difficile d'estimer des modèles de régression moyenne mobile que d'estimer des modèles à aléas autorégressifs. Pour en comprendre la raison, supposons que l'on veuille estimer le modèle

$$y_t = x_t(\beta) + u_t, \quad u_t = \varepsilon_t - \alpha\varepsilon_{t-1}, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.61)$$

Par rapport à (10.57), nous avons ignoré l'indice de  $\alpha$  et changé son signe par commodité; le changement de signe est bien sûr une pure normalisation. Faisons l'hypothèse asymptotiquement anodine de la nullité de l'innovation non observée  $\varepsilon_0$  (nous discuterons des techniques qui ne font pas usage de cette hypothèse plus loin). Alors nous voyons que

$$\begin{aligned} y_1 &= x_1(\beta) + \varepsilon_1 \\ y_2 &= x_2(\beta) - \alpha(y_1 - x_1(\beta)) + \varepsilon_2 \\ y_3 &= x_3(\beta) - \alpha(y_2 - x_2(\beta)) - \alpha^2(y_1 - x_1(\beta)) + \varepsilon_3, \end{aligned} \quad (10.62)$$

et ainsi de suite. En posant les définitions

$$\begin{aligned} y_0^* &= 0; \quad y_t^* = y_t + \alpha y_{t-1}^*, \quad t = 1, \dots, n; \\ x_0^* &= 0; \quad x_t^*(\beta, \alpha) = x_t(\beta) + \alpha x_{t-1}^*(\beta, \alpha), \quad t = 1, \dots, n, \end{aligned} \quad (10.63)$$

on peut écrire les équations (10.62) comme

$$y_t = -\alpha y_{t-1}^* + x_t^*(\beta, \alpha) + \varepsilon_t, \quad (10.64)$$

qui atteste clairement que nous devons gérer un modèle de régression non linéaire. Mais la fonction de régression dépend de l'échantillon tout entier

jusqu'à la période  $t$ , puisque  $y_{t-1}^*$  dépend de toutes les valeurs précédentes de  $y_t$  et que  $x_t^*$  dépend de  $x_{t-i}(\beta)$  pour tout  $i \geq 0$ . Dans le cas nullement improbable où  $\alpha = 1$ , la dépendance de  $y_t$  aux valeurs passées ne tend même pas à diminuer au fur et à mesure que l'on recule dans le temps. Si l'on dispose d'un programme ingénieux de moindres carrés non linéaires nous permettant de définir la fonction de régression de façon récursive, comme dans (10.63), l'estimation de (10.64) n'est pas plus difficile que l'estimation d'autres modèles de régression non linéaire. Mais si l'on ne dispose pas de tels logiciels, cette phase d'estimation peut être assez délicate.

En rejetant l'hypothèse de la distribution normale des aléas, le modèle (10.61) devient

$$y_t = x_t(\beta) + u_t, \quad u_t = \varepsilon_t - \alpha\varepsilon_{t-1}, \quad \varepsilon_t \sim \text{NID}(0, \omega^2). \quad (10.65)$$

Nous avons fait l'hypothèse asymptotiquement banale de nullité de l'innovation non observée  $\varepsilon_0$ . Bien que banale asymptotiquement, cette hypothèse est évidemment erronée, puisqu'en accord avec (10.65)  $\varepsilon_0$  doit être distribuée selon la  $N(0, \omega^2)$ . La façon la plus simple de prendre en considération ce fait a été suggérée par MacDonald et MacKinnon (1985); notre traitement suit le leur.

La fonction de logvraisemblance concentrée pour le modèle (10.65) est

$$C - \frac{n}{2} \log \left( (\mathbf{y} - \mathbf{x}(\beta))^\top \mathbf{\Delta}^{-1}(\alpha) (\mathbf{y} - \mathbf{x}(\beta)) \right) - \frac{1}{2} \log |\mathbf{\Delta}(\alpha)|, \quad (10.66)$$

où  $\omega^2 \mathbf{\Delta}(\alpha)$  est la matrice de covariance du vecteur  $\mathbf{u}$  des aléas, donnée par l'expression (10.60).<sup>4</sup> Comme Box et Jenkins (1976) et d'autres en ont discuté, le terme Jacobien  $\frac{1}{2} \log |\mathbf{\Delta}(\alpha)|$  est

$$\frac{1}{2} \log(1 - \alpha^2) - \frac{1}{2} \log(1 - \alpha^{2n+2}). \quad (10.67)$$

Lorsque  $|\alpha| = 1$ , les deux termes de (10.67) ne sont pas définis. Dans cette situation, on peut montrer, grâce à la règle de l'Hôpital, que

$$\lim_{|\alpha| \rightarrow 1} \left( \frac{1}{2} \log(1 - \alpha^2) - \frac{1}{2} \log(1 - \alpha^{2n+2}) \right) = -\frac{1}{2} \log(n+1).$$

Ce résultat autorise l'évaluation de la fonction de logvraisemblance (10.66) pour toute valeur de  $\alpha$  comprise dans la région d'inversibilité  $-1 \leq \alpha \leq 1$ .

<sup>4</sup> En réalité, l'expression (10.66) pourrait correspondre à la fonction de logvraisemblance concentrée pour un modèle de régression non linéaire dont les aléas suivent un processus quelconque **autorégressif moyenne mobile**, ou processus **ARMA**, à condition que  $\mathbf{\Delta}(\alpha)$  soit remplacée par la matrice de covariance du vecteur  $\mathbf{u}$  impliqué dans le processus ARMA.



Il est important d'être à même de gérer le cas où  $|\alpha| = 1$ , puisqu'en pratique on obtient assez souvent des estimations ML avec  $|\hat{\alpha}| = 1$ , en particulier lorsque la taille de l'échantillon est faible; voir, par exemple, Osborn (1976) et Davidson (1981). La raison est que si l'on concentre la fonction de logvraisemblance par rapport à  $\beta$  et à  $\omega$  afin d'obtenir  $\ell^c(\alpha)$ , on trouvera que  $\ell^c(\alpha)$  a la même valeur pour  $\alpha$  et pour  $1/\alpha$ . Bien évidemment, cela justifie l'imposition de la condition d'inversibilité selon laquelle  $|\alpha| \leq 1$ . Ainsi, si  $\ell^c(\alpha)$  croît quand  $\alpha \rightarrow 1$  ou quand  $\alpha \rightarrow -1$ , elle doit avoir un maximum précisément en  $\alpha = 1$  ou  $\alpha = -1$ . Cela manifeste un trait caractéristique peu désirable du modèle (10.65). Lorsque  $|\hat{\alpha}| = 1$ , il est impossible de faire des inférences sur  $\alpha$  de la façon habituelle, puisqu'alors  $\hat{\alpha}$  est sur la frontière de l'espace paramétrique. Puisque  $\hat{\alpha}$  peut être égale à  $\pm 1$  avec une probabilité finie, l'usage de la distribution normale pour approximer la distribution avec un échantillon fini est une procédure quelque peu douteuse. Ainsi, si  $\hat{\alpha}$  est égale à 1 ou même proche de 1 en valeur absolue, il faudrait prendre beaucoup de précautions en pratiquant des inférences sur  $\alpha$ . Bien sûr, lorsque  $n \rightarrow \infty$ , la convergence de  $\hat{\alpha}$  signifie que le nombre de fois où  $|\hat{\alpha}| = 1$  tend vers zéro, sauf si  $|\alpha_0| = 1$ .

L'évaluation directe de (10.66) n'est pas aisée; consulter Pesaran (1973), Osborn (1976), et Balestra (1980), parmi d'autres.<sup>5</sup> Nous allons donc ruser et fournir un moyen alternatif pour y parvenir. Souvenons-nous des équations (10.62), dans lesquelles nous avons écrit de façon explicite  $y_1, \dots, y_n$  comme des fonctions des valeurs retardées de  $x_t(\beta)$  et des valeurs retardées de  $y_t$ . Nous avons la possibilité de récrire ces équations en tenant compte de l'observation numéro zéro, comme

$$\begin{aligned} 0 &= -v + \varepsilon_0 \\ y_1 &= x_1(\beta) - \alpha v + \varepsilon_1 \\ y_2 &= x_2(\beta) - \alpha(y_1 - x_1(\beta)) - \alpha^2 v + \varepsilon_2 \\ y_3 &= x_3(\beta) - \alpha(y_2 - x_2(\beta)) - \alpha^2(y_1 - x_1(\beta)) - \alpha^3 v + \varepsilon_3, \end{aligned} \tag{10.68}$$

et ainsi de suite. Nous avons ajouté ici à la fois une observation et un paramètre aux équations (10.62). L'observation supplémentaire est l'observation zéro qui, telle qu'elle est ici écrite, exprime simplement l'égalité par *définition* entre le paramètre inconnu  $v$  et l'aléa  $\varepsilon_0$ . Ce paramètre inconnu apparaît aussi dans toutes les observations suivantes, multiplié par des puissances toujours plus fortes de  $\alpha$ , de façon à refléter la dépendance de  $y_t$  à  $\varepsilon_0$  pour toutes les observations. Notons que parce que nous avons ajouté une observation et un paramètre, nous n'avons pas modifié le nombre de degrés de liberté (c'est-à-dire le nombre d'observations moins le nombre de paramètres estimés).

<sup>5</sup> Une toute autre approche concernant l'estimation des modèles à aléas obéissant à un processus moyenne mobile fut proposée par Harvey et Phillips (1979) et par Gardner, Harvey, et Phillips (1980). Elle nécessite un logiciel spécialisé.

Si nous adoptons les définitions

$$\begin{aligned} y_0^* &= 0; & y_t^* &= y_t + \alpha y_{t-1}^*, \quad t = 1, \dots, n; \\ x_0^* &= 0; & x_t^*(\beta, \alpha) &= x_t(\beta) + \alpha x_{t-1}^*(\beta, \alpha), \quad t = 1, \dots, n; \\ z_0^* &= -1; & z_t^* &= \alpha z_{t-1}^*, \end{aligned}$$

nous pouvons écrire les équations (10.68) sous la forme

$$y_t^*(\alpha) = x_t^*(\beta, \alpha) + v z_t^* + \varepsilon_t, \quad (10.69)$$

les rendant très comparables à un modèle de régression non linéaire. La somme des résidus au carré serait alors

$$\sum_{t=1}^n (y_t^*(\alpha) - x_t^*(\beta, \alpha) - v z_t^*)^2. \quad (10.70)$$

Lorsqu'elle est évaluée en la valeur de  $v$  qui la minimise, la somme des résidus au carré (10.70) est égale à la somme des carrés généralisée

$$(\mathbf{y} - \mathbf{x}(\beta))^\top \mathbf{\Delta}^{-1}(\alpha) (\mathbf{y} - \mathbf{x}(\beta)), \quad (10.71)$$

qui apparaît dans la fonction de logvraisemblance (10.66); on doit la démonstration de ce résultat à Pagan et Nicholls (1976). Nous pouvons donc remplacer (10.71) par (10.70) à l'intérieur de (10.66), ce qui la rend beaucoup plus simple à évaluer. Lorsque  $x_t(\beta)$  est linéaire, l'approche la plus simple consiste probablement à explorer  $\alpha$  dans l'intervalle allant de  $-1$  à  $+1$ , puisqu'on peut alors minimiser la SSR (10.70) par OLS et connecter le résultat sur (10.66) afin d'évaluer la fonction de logvraisemblance. Lorsque  $x_t(\beta)$  est non linéaire on peut directement maximiser (10.66) par rapport à  $\alpha$  et  $\beta$  simultanément. Si  $x_t(\beta)$  est linéaire et qu'il n'y a pas de variable dépendante retardée parmi les régresseurs, on pourra pratiquer des inférences sur  $\beta$  en utilisant la matrice de covariance des OLS habituelle de  $\hat{\beta}$  conditionnellement à  $\hat{\alpha}$  provenant de (10.69). Autrement, on peut utiliser la régression de Gauss-Newton correspondant à (10.69).

Il est possible de spécifier un modèle qui combine les composantes autorégressive et moyenne mobile. Le résultat est le modèle **ARMA**( $p, q$ ),

$$A(L, \rho) u_t = B(L, \alpha) \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.72)$$

Le membre de gauche de (10.72) ressemble au modèle AR( $p$ ), et celui de droite ressemble au modèle MA( $q$ ) (10.58). L'avantage des modèles ARMA est qu'un modèle relativement économe, tel que ARMA(1, 1) ou ARMA(2, 1), offre souvent une représentation d'une série temporelle aussi bonne que celle obtenue avec un modèle AR ou MA qui serait plus coûteux à gérer.

Enfin, il nous faut mentionner la classe des **modèles ARIMA**. Ceux-ci sont simplement des modèles ARMA appliqués à des données qui ont été différenciées un certain nombre (entier) de fois, disons  $d$ . Ainsi le modèle **ARIMA**( $p, d, q$ ) est

$$A(L, \boldsymbol{\rho})(1 - L)^d u_t = B(L, \boldsymbol{\alpha}) \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.73)$$

Lorsque  $d = 0$ , il se ramène à un modèle ARMA( $p, q$ ) standard. Le I dans ARIMA signifie **intégré**, puisqu'une série intégrée doit être différenciée pour être stationnaire. La différenciation est souvent utilisée pour provoquer la stationnarité dans les séries temporelles qui autrement ne seraient pas stationnaires. Bien que nous ne nous attendons pas à ce que les aléas qu'on associe à un modèle de régression ne soient pas stationnaires, de nombreuses séries temporelles économiques sont elles-mêmes (en apparence) non stationnaires et devraient normalement être différenciées avant d'être utilisées dans un modèle économétrique. Nous discuterons des séries temporelles non stationnaires au cours des Chapitres 19 et 20.

Notre traitement des modèles de régression à aléas MA fut bref et restreint au cas MA(1). Ceux qui ont besoin d'estimer de tels modèles, ou des modèles à aléas ARMA, sont généralement informés qu'il faut utiliser un logiciel spécialisé, qui emploiera de façon typique des techniques d'estimation telles que celles discutées par Newbold (1974), Box et Jenkins (1976), Dent (1977), Ansley (1979), Zinde-Walsh et Galbraith (1991), et Galbraith et Zinde-Walsh (1992).

## 10.8 TEST D'AUTOCORRÉLATION

Une fraction conséquente de toute la littérature en économétrie a été consacrée au problème des tests d'autocorrélation des aléas des modèles de régression. La plus grande part de cette fraction a traité les tests de l'hypothèse nulle selon laquelle les aléas d'un modèle de régression linéaire sont indépendants entre eux contre l'hypothèse alternative selon laquelle ils obéissent à un processus AR(1). Bien que l'autocorrélation soit certainement un phénomène très répandu avec les données temporelles et que par conséquent son test soit à l'évidence important, la masse d'efforts déployée pour traiter ce problème paraît quelque peu disproportionnée. Comme nous allons le voir, les tests asymptotiques pour l'autocorrélation peuvent être directement dérivés comme des applications de la régression de Gauss-Newton. A l'exception du cas où il est possible de pratiquer des inférences exactes avec des échantillons finis, il n'y a aucune raison d'employer des procédures plus spécialisées et difficiles.

Supposons que l'on désire tester l'hypothèse nulle selon laquelle les erreurs  $u_t$  du modèle

$$y_t = x_t(\boldsymbol{\beta}) + u_t \quad (10.74)$$

sont indépendantes entre elles contre l'hypothèse alternative selon laquelle elles obéissent à un processus AR(1). Comme nous l'avons déjà vu, pour les observations  $t = 2, \dots, n$ , ce modèle alternatif peut s'écrire comme

$$y_t = x'_t(\boldsymbol{\beta}, \rho) + \varepsilon_t \equiv x_t(\boldsymbol{\beta}) + \rho(y_{t-1} - x_{t-1}(\boldsymbol{\beta})) + \varepsilon_t, \quad (10.75)$$

où  $\varepsilon_t$  est supposé être IID(0,  $\omega^2$ ). Ainsi que nous l'avons vu au cours du Chapitre 6, tout ensemble de contraintes sur les paramètres d'une fonction de régression non linéaire peut être testé en exécutant une régression de Gauss-Newton évaluée avec les estimations convergentes au taux  $n^{1/2}$  sous l'hypothèse nulle. Ces estimations seraient typiquement des estimations NLS contraintes, mais pas nécessairement. Ainsi, dans ce cas, on peut tester  $\rho = 0$  en régressant  $y_t - x'_t$  sur les dérivées de la fonction de régression  $x'_t(\boldsymbol{\beta}, \rho)$  par rapport à tous ses paramètres, sachant qu'autant  $x'_t$  que ses dérivées seront évaluées avec les estimations du vecteur de paramètres  $[\boldsymbol{\beta} : \rho]$  sous l'hypothèse nulle. En supposant que (10.74) a été estimée par moindres carrés, ces estimations sont simplement  $[\tilde{\boldsymbol{\beta}} : 0]$ , où  $\tilde{\boldsymbol{\beta}}$  désigne l'estimation par moindres carrés de  $\boldsymbol{\beta}$  conditionnellement à  $\rho = 0$ .<sup>6</sup> Puisque les dérivées sont

$$\frac{\partial x'_t}{\partial \beta_i} = \mathbf{X}_t(\boldsymbol{\beta}) - \rho \mathbf{X}_{t-1}(\boldsymbol{\beta}); \quad \frac{\partial x'_t}{\partial \rho} = y_{t-1} - x_{t-1}(\boldsymbol{\beta}),$$

la GNR appropriée est

$$y_t - x_t(\tilde{\boldsymbol{\beta}}) = \mathbf{X}_t(\tilde{\boldsymbol{\beta}})\mathbf{b} + r(y_{t-1} - x_{t-1}(\tilde{\boldsymbol{\beta}})) + \text{résidu}$$

que l'on peut écrire de façon compacte comme

$$\tilde{\mathbf{u}} = \tilde{\mathbf{X}}\mathbf{b} + r\tilde{\mathbf{u}}_{-1} + \text{résidus}, \quad (10.76)$$

où  $\tilde{\mathbf{u}}$  désigne le vecteur des résidus des moindres carrés dont l'élément type est  $y_t - x_t(\tilde{\boldsymbol{\beta}})$ ,  $\tilde{\mathbf{X}}$  désigne la matrice des dérivées de la fonction de régression  $x_t(\boldsymbol{\beta})$  dont l'élément type est  $X_{ti}(\tilde{\boldsymbol{\beta}})$ , et où  $\tilde{\mathbf{u}}_{-1}$  désigne le vecteur dont l'élément type est  $\tilde{u}_{t-1}$ . C'est une régression extrêmement simple à mettre en œuvre, en particulier lorsque le modèle originel (10.74) est linéaire. Dans ce contexte, puisque  $\tilde{\mathbf{X}}$  est tout simplement la matrice des régresseurs, il suffit de régresser les résidus sur les régresseurs originels et sur les résidus retardés. La statistique de test pourrait être soit  $n$  fois le  $R^2$  non centré soit le  $t$  de Student ordinaire pour  $r = 0$ . La première sera asymptotiquement distribuée selon la  $\chi^2(1)$  sous l'hypothèse nulle, la seconde sera  $N(0, 1)$ . Dans la pratique il est généralement préférable d'utiliser le  $t$  de Student et de le confronter à la distribution de Student aux degrés de liberté convenables; consulter Kiviet (1986).

<sup>6</sup> Il existe un résultat concernant l'échantillon à utiliser,  $t = 1$  à  $n$  ou  $t = 2$  à  $n$ , pour estimer  $\tilde{\boldsymbol{\beta}}$ ; nous en discuterons plus loin.

La discussion qui précède a fait l'impasse sur le problème pratique de la gestion de l'observation initiale. Le modèle alternatif (10.75) est défini uniquement pour les observations allant de 2 à  $n$ , ce qui suggère que  $\tilde{\beta}$  devrait également être obtenu par une estimation reposant sur un échantillon temporel plus court. Par chance, cela est totalement inutile. Une première approche consiste à exécuter une GNR valable pour les observations allant de 2 à  $n$  uniquement. Le seul problème de cette approche est que  $\tilde{u}$  ne sera plus orthogonal à  $\tilde{X}$ . Par conséquent, le  $R^2$  pour la GNR ne sera pas nul même si  $\tilde{u}_{-1}$  n'est pas pris en compte, et il en découle que la version  $nR^2$  des tests basés sur cette régression peut avoir tendance à rejeter l'hypothèse nulle trop fréquemment avec des échantillons finis. Cela ne posera pas de difficulté si la statistique de test est le  $t$  de Student pour  $r = 0$ . Une seconde approche consiste à obtenir  $\tilde{\beta}$  à partir d'une estimation opérant sur l'échantillon complet et exécuter une GNR sur la période couverte par l'échantillon entier, en initialisant le  $\tilde{u}_0$  non observé à zéro.

Lorsque le modèle originel est linéaire, une légère modification de cette procédure est envisageable. Parce que  $X\tilde{\beta}$  se situe dans  $\mathcal{S}(X)$ , la régression

$$y = Xc + r\tilde{u}_{-1} + \text{résidus} \quad (10.77)$$

aura dans ce cas exactement la même somme des résidus au carré, et exactement le même  $t$  de Student pour  $r = 0$ , que la régression de test d'origine (10.76). Ainsi, pour des modèles linéaires, le moyen le plus aisé de tester des aléas AR(1) est simplement d'exécuter à nouveau la régression originelle avec un régresseur additionnel, égal à 0 pour l'observation 1 et égal à  $\tilde{u}_{t-1}$  pour les observations qui suivent. Il faut ensuite utiliser le  $t$  de Student habituel pour ce régresseur supplémentaire, puisqu'à l'évidence l'usage du  $nR^2$  à partir de (10.77) *n'est pas* valable.

L'extension de ces procédures aux tests des aléas AR de plus haut degré est immédiate. Supposons que l'hypothèse alternative soit telle que les  $u_t$  de (10.74) suivent un processus d'aléas AR( $p$ ). Le modèle alternatif peut s'écrire

$$y_t = x_t(\beta) + \sum_{j=1}^p \rho_j (y_{t-j} - x_{t-j}(\beta)) + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2),$$

ce qui implique que la régression de test analogue à (10.76) est

$$\tilde{u} = \tilde{X}b + \sum_{j=1}^p r_j \tilde{u}_{-j} + \text{résidus}. \quad (10.78)$$

Une statistique de test envisageable est  $n$  fois le  $R^2$  non centré de cette régression, asymptotiquement distribué suivant une  $\chi^2(p)$  sous l'hypothèse nulle. Une autre statistique de test, qui possède probablement de meilleures propriétés avec des échantillons finis, est le  $F$  de Fisher asymptotique pour

$r_1 = r_2 = \dots = r_p = 0$ . Cette statistique de test aura  $p$  et  $n - k - p$  degrés de liberté, sous la condition que (10.78) soit exécutée sur l'échantillon tout entier, en utilisant des zéros pour compléter les éléments initiaux de  $\tilde{\mathbf{u}}_{-j}$ . Lorsque le modèle originel est linéaire, il est toujours valable de remplacer la régressande  $\tilde{\mathbf{u}}$  par la variable originelle dépendante  $\mathbf{y}$ , comme dans (10.77). Lorsque cette phase est effectuée, la variante  $nR^2$  du test ne peut bien sûr plus être utilisée.

Supposons que l'on veuille tester l'hypothèse nulle contre une spécification des aléas MA(1) plutôt que AR(1). Le modèle alternatif serait alors le modèle plutôt compliqué donné par (10.62) ou (10.64). Les dérivées de ce modèle par rapport à  $\beta$  et  $\alpha$  sont également assez compliquées, mais elles se simplifient considérablement lorsqu'elles sont évaluées sous l'hypothèse nulle  $\alpha = 0$ . En réalité, lorsque l'on évalue ces dérivées en  $[\tilde{\beta} : 0]$ , nous voyons que, pour toutes les observations, la dérivée par rapport à  $\beta_i$  est  $X_{ti}(\tilde{\beta})$  et, pour les observations allant de 2 à  $n$ , la dérivée par rapport à  $\alpha$  est  $y_{t-1} - x_{t-1}(\tilde{\beta})$ .<sup>7</sup> Ainsi la GNR qui permet le test contre la spécification des aléas MA(1) est *identique* à celle du test contre la spécification AR(1). Ceci est une conséquence du fait que, sous l'hypothèse nulle d'absence d'autocorrélation, les modèles de régression avec des aléas AR(1) et MA(1) sont ce que Godfrey et Wickens (1982) appellent des **alternatives localement équivalentes**, c'est-à-dire des modèles qui ont des dérivées identiques lorsqu'elles sont évaluées sous l'hypothèse nulle. Puisque les tests basés sur la GNR utilisent uniquement les informations relatives aux dérivées premières du modèle alternatif, il n'est pas surprenant que, si deux modèles sont localement équivalents dans ce sens sous une certaine hypothèse nulle, les GNR qui en résultent soient identiques; voir Godfrey (1981).

Pour voir qu'un processus AR(1) est localement équivalent à un processus MA(1), souvenons-nous que le premier processus peut être écrit comme

$$u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \dots$$

Si l'on différentie le membre de droite par rapport à  $\rho$  et si l'on évalue ces dérivées en  $\rho = 0$ , le résultat est tout simplement  $\varepsilon_{t-1}$ . Mais cela est également la dérivée du processus MA(1) (10.57) par rapport à son unique paramètre. Ainsi nous voyons que les processus AR(1) et MA(1) sont en réalité des modèles alternatifs localement équivalents.

Prenant en considération le résultat selon lequel la même régression de Gauss-Newton peut être employée pour tester le modèle originel contre la spécification MA(1) aussi bien que celle AR(1), il ne devrait pas être surprenant de savoir que la GNR qui teste contre la spécification MA( $q$ ) des aléas est identique à celle qui teste contre la spécification AR( $q$ ). Peut-être est-il plus surprenant que la même régression artificielle se révèle adéquate

<sup>7</sup> Puisque pour l'observation 1 cette dérivée est nulle, notre suggestion d'utiliser un zéro au lieu de l'aléa non connu  $\tilde{u}_0$  est tout à fait appropriée ici.

pour tester contre une spécification  $\text{ARMA}(p, q)$  des aléas, avec  $p + q$  retards de  $\tilde{\mathbf{u}}$  compris dans la régression. Pour plus de détails, voir Godfrey (1978b, 1988).

L'usage d'un outil comparable à la régression de Gauss-Newton pour tester l'autocorrélation fut initialement suggérée par Durbin (1970) dans un article qui introduisait ce qui est connu comme le **test en  $h$  de Durbin**. Cette dernière procédure, que nous ne détaillons pas, est un test asymptotique d'aléas  $\text{AR}(1)$  que l'on peut utiliser lorsque l'hypothèse nulle est un modèle de régression linéaire qui comprend la variable dépendante retardée une fois, et éventuellement davantage, dans les régresseurs. Le test en  $h$  peut se calculer avec une simple calculatrice de poche à partir des résultats sur la régression d'origine retournés par la plupart des progiciels de régression, bien que quelquefois il soit impossible de le calculer parce qu'il serait nécessaire de trouver la racine carrée d'un nombre négatif. Pour des raisons qui nous paraissent difficiles à admettre aujourd'hui (mais qui sont sans doute liées à l'état peu avancé du matériel informatique et des logiciels d'économétrie dans les années 70), le test en  $h$  de Durbin fut largement utilisé depuis, alors que ce que l'on appelle sa **procédure alternative**, un test en  $t$  basé sur la GNR modifiée (10.77), était parfaitement ignorée durant un certain temps.<sup>7</sup> Il fut finalement redécouvert et étendu par Breusch (1978) et Godfrey (1978a, 1978b). Tous ces articles supposaient que les aléas étaient normalement distribués, et ils présentaient des tests basés sur la GNR comme des tests du multiplicateur de Lagrange basés sur l'estimation par maximum de vraisemblance. Bien sûr, cette hypothèse de normalité n'est pas du tout nécessaire.

De même, toute hypothèse relative à la présence, ou à l'absence, de variables dépendantes retardées dans la fonction de régression  $x_t(\beta)$  est inutile. Tout ce que nous demandons, c'est que cette fonction satisfasse les conditions de régularité du Chapitre 5, de manière à ce que les estimations par moindres carrés non linéaires soient convergentes et asymptotiquement normales à la fois sous l'hypothèse nulle et sous l'hypothèse alternative. Comme ce qui précède l'implique, et comme nous en discuterons plus loin, de nombreux tests d'autocorrélation nécessitent que  $x_t(\beta)$  ne dépende pas de variable dépendante retardée, et toute la littérature citée dans le paragraphe précédent fut écrite dans le but spécifique de manipuler le cas dans lequel  $x_t(\beta)$  est linéaire et dépend d'une, ou de plusieurs valeurs retardées de la variable dépendante.

Le problème avec les tests fondés sur la GNR est qu'ils ne sont valables qu'asymptotiquement. Ceci est vrai que  $x_t(\beta)$  soit linéaire ou pas, parce que  $\tilde{\mathbf{u}}_{-1}$  n'est qu'une estimation de  $\mathbf{u}_{-1}$ . En fait, comme nous l'avons vu

<sup>7</sup> Maddala et Rao (1973), Spencer (1975), et Inder (1984), parmi d'autres auteurs, ont fourni une preuve grâce aux expériences Monte Carlo du test en  $h$  de Durbin en le comparant au test basé sur la GNR. Cette preuve ne suggère aucune raison sérieuse de préférer l'un à l'autre. Ainsi l'aspect plus pratique et la polyvalence plus large du test basé sur la GNR sont sans doute les facteurs principaux qui jouent en sa faveur.

au cours de la Section 5.6,  $\tilde{\mathbf{u}} \stackrel{a}{=} \mathbf{M}_0 \mathbf{u}$ , où  $\mathbf{M}_0 \equiv \mathbf{I} - \mathbf{X}_0(\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top$  et  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$ . Cela correspond simplement à l'égalité asymptotique (5.57). L'égalité asymptotique est remplacée par une égalité exacte si  $\mathbf{x}(\beta) = \mathbf{X}\beta$ . Cette relation montre clairement que même si les  $u_t$  sont indépendants entre eux, les  $\tilde{u}_t$  ne le seront pas et le test peut alors rejeter à tort l'hypothèse nulle. Le problème s'estompe lorsque la taille de l'échantillon est suffisamment importante, puisque le vecteur  $\tilde{\mathbf{u}}$  tend vers  $\mathbf{u}$  lorsque la taille de l'échantillon tend vers l'infini, à condition bien sûr que le modèle estimé contienne le DGP. Dans la pratique, cela ne semble pas être un problème délicat même si la taille de l'échantillon est modeste (disons 50 ou plus), à condition d'employer la configuration correcte du test. Les résultats de Kiviet (1986) suggèrent que les tests en  $F$  basés sur la GNR fonctionnent généralement assez bien même si l'échantillon contient une vingtaine d'observations (à condition que le nombre de régresseurs soit également petit), mais ils suggèrent aussi que les tests calculés comme le  $nR^2$  sont moins fiables et peuvent tendre à rejeter trop souvent l'hypothèse nulle lorsqu'il y a effectivement une autocorrélation.

Le test d'autocorrélation le plus populaire en économétrie est conçu pour manipuler les problèmes qui résultent du fait que  $\tilde{\mathbf{u}}$  ne possède pas vraiment les mêmes propriétés que  $\mathbf{u}$ , mais uniquement pour les modèles linéaires sans variable dépendante retardée, et avec des aléas supposés suivre une distribution normale. C'est la **statistique  $d$**  proposée par Durbin et Watson (1950, 1951) auquel on se réfère en tant que **statistique DW**. La définition de cette statistique est

$$d = \frac{\sum_{t=2}^n (\tilde{u}_t - \tilde{u}_{t-1})^2}{\sum_{t=1}^n \tilde{u}_t^2}, \quad (10.79)$$

où, comme d'habitude,  $\tilde{u}_t$  est le résidu  $t$  de l'estimation OLS de la régression que l'on teste pour une autocorrélation éventuelle à l'ordre un. Cette régression peut être linéaire ou non linéaire, bien que les résultats avec des échantillons finis dépendent de la linéarité.

Il est aisé de voir que le numérateur du  $d$  de Durbin est approximativement égal à

$$2 \left( \sum_{t=2}^n \tilde{u}_t^2 - \sum_{t=2}^n \tilde{u}_t \tilde{u}_{t-1} \right). \quad (10.80)$$

Ainsi le  $d$  de Durbin lui-même est sensiblement égal à  $2 - 2\tilde{\rho}$ , où  $\tilde{\rho}$  est l'estimation de  $\rho$  obtenue en régressant  $\tilde{u}_t$  sur  $\tilde{u}_{t-1}$ :

$$\tilde{\rho} = \frac{\sum_{t=2}^n \tilde{u}_t \tilde{u}_{t-1}}{\sum_{t=2}^n \tilde{u}_{t-1}^2}. \quad (10.81)$$

Ces résultats ne sont vrais qu'en tant qu'approximations parce que (10.79), (10.80) et (10.81) traitent la première et la dernière observations différemment.



N'importe quel effet dû à ces observations doit, cependant, disparaître asymptotiquement. Ainsi il est clair qu'avec des échantillons de taille raisonnable, la statistique  $d$  doit varier entre 0 et 4, et qu'une valeur de 2 correspond à une totale absence d'autocorrélation. Les valeurs de la statistique  $d$  inférieures à 2 correspondent à  $\tilde{\rho} > 0$ , alors que des valeurs supérieures à 2 correspondent à  $\tilde{\rho} < 0$ .

Il est possible, mais embarrassant d'un point de vue calculatoire, de calculer la distribution exacte de la statistique  $d$  lorsque les  $u_t$  sont normaux, lorsque le modèle de régression sous-jacent est linéaire, et lorsque  $\mathbf{X}$  ne possède que des régresseurs fixes. Cette distribution dépend forcément de  $\mathbf{X}$ . Les calculs utilisent le fait que la statistique  $d$  peut s'écrire comme

$$\frac{\mathbf{u}^\top \mathbf{M}_X \mathbf{A} \mathbf{M}_X \mathbf{u}}{\mathbf{u}^\top \mathbf{M}_X \mathbf{u}}, \quad (10.82)$$

où  $\mathbf{A}$  est la matrice de dimension  $n \times n$

$$\begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

A partir de (10.82), la statistique  $d$  est considérée comme un rapport de deux formes quadratiques de variables aléatoires normales, et la distribution de telles quantités peut être évaluée en utilisant de nombreuses techniques différentes; voir Durbin et Watson (1971) et Savin et White (1977).

La plupart des utilisateurs n'essayent jamais de calculer la distribution exacte de la statistique  $d$  correspondant à leur matrice  $\mathbf{X}$ . Au lieu de cela, ils utilisent la propriété qui veut que les valeurs critiques de sa distribution sont comprises entre deux bornes,  $d_L$  et  $d_U$ , qui dépendent de la taille de l'échantillon,  $n$ , du nombre de régresseurs,  $k$ , et de la présence ou de l'absence d'un terme constant. On trouvera les tables de  $d_L$  et  $d_U$  dans les ouvrages d'économétrie et dans des articles tels que ceux de Durbin et Watson (1951) et de Savin et White (1977). A titre d'exemple, lorsque  $n = 50$  et  $k = 6$  (en comptabilisant la constante comme un régresseur), pour un test contre  $\rho > 0$  au niveau .05,  $d_L = 1.335$  et  $d_U = 1.771$ . Ainsi, si quelqu'un calculait une statistique  $d$  pour cette taille d'échantillon et ce nombre de régresseurs et que sa valeur était inférieure à 1.335, il pourrait décider à raison de rejeter l'hypothèse nulle d'absence d'autocorrélation au niveau .05. Si la valeur de la statistique était supérieure à 1.771, il pourrait décider de ne pas la rejeter. Cependant, si la valeur de cette statistique se situait dans la "région d'indécision" entre 1.335 et 1.771, il ne serait fixé ni sur son rejet ni sur

son acceptation. Lorsque la taille d'échantillon est petite, et tout particulièrement lorsqu'elle est petite relativement au nombre de régresseurs, la région d'indécision peut être très étendue. Cela signifie que la statistique  $d$  peut ne pas porter beaucoup d'information lorsqu'elle est utilisée conjointement aux tables de  $d_L$  et  $d_U$ .<sup>8</sup> Dans de telles circonstances, il faudrait sans doute calculer la distribution exacte de la statistique si l'on désire procéder à des inférences à partir de la statistique  $d$  avec un échantillon faible. Un petit nombre de progiciels, tel que SHAZAM, le permet. À l'évidence, parce que les hypothèses d'erreurs normales et de régresseurs fixes sont trop fortes même les inférences "exactes" avec un échantillon fini ne sont en réalité que des approximations.

Comme nous l'avons déjà mentionné, la statistique  $d$  n'est pas valable, même asymptotiquement, lorsque  $\mathbf{X}$  comprend des valeurs retardées de la variable dépendante. La meilleure manière de comprendre pourquoi c'est le cas est d'utiliser le fait que la statistique  $d$  est asymptotiquement équivalente au  $t$  de Student de l'estimation de  $\rho$  dans la régression

$$\tilde{\mathbf{u}} = \rho \tilde{\mathbf{u}}_{-1} + \text{résidus}; \quad (10.83)$$

voir la discussion qui mène à (10.81). La seule différence qui réside entre (10.83) et la régression de Gauss-Newton (10.76), qui génère un  $t$  de Student asymptotiquement valable pour le coefficient de  $\tilde{\mathbf{u}}_{-1}$ , est que (10.83) ne possède pas de matrice  $\tilde{\mathbf{X}}$  parmi les régresseurs. Nous avons discuté de ce détail dans la Section 6.4, où nous avons vu que le  $t$  de Student correct, à partir de (10.76), doit être asymptotiquement supérieur (ou du moins pas inférieur) à celui généralement incorrect, donné par (10.83).

Si  $\mathbf{x}(\beta)$  ne dépend pas de variables dépendantes retardées,  $x_t(\beta)$  et par conséquent  $\mathbf{X}_t(\beta)$  doit être non corrélé avec toutes les valeurs retardées de  $u_t$ . Par la suite,  $\tilde{\mathbf{X}}$  n'aura pas de pouvoir explicatif sur  $\tilde{\mathbf{u}}_{-1}$ , asymptotiquement, et les  $t$  de Student donnés par (10.76) et par (10.83) sont asymptotiquement identiques. Mais si  $\mathbf{x}(\beta)$  dépend de variables dépendantes retardées,  $\mathbf{X}_t(\beta)$  sera corrélé avec certaines valeurs de  $u_t$ , puisque des valeurs retardées de la variable dépendante sont sûrement corrélées avec des valeurs retardées des aléas. Alors le  $t$  de Student de (10.83) sera asymptotiquement plus faible que celui de (10.76), et la statistique  $d$  sera par conséquent biaisée vers 2. Cependant, elle peut quand même porter de l'information. Si sa valeur était telle qu'elle nous permette de rejeter l'hypothèse nulle d'absence d'autocorrélation si  $\mathbf{x}(\beta)$  ne dépendait pas de variables dépendantes retardées, alors une statistique de test correcte basée sur la GNR nous permettrait certainement de le faire.

<sup>8</sup> Il y a des raisons de croire que lorsque les régresseurs changent légèrement, un cas de figure qui pourrait être très commun avec des données chronologiques,  $d_U$  fournit une meilleure approximation que  $d_L$ . Voir Hannan et Terrell (1966).

Nous terminons cette section par une brève discussion sur d'autres tests d'autocorrélation. Kobayashi (1991) proposa un test qui est exact à l'ordre  $n^{-1/2}$  pour les modèles de régression non linéaire sans variable dépendante retardée. Il est basé sur l'estimation de  $\rho$  sous l'hypothèse alternative, qui est alors corrigée pour atténuer le biais. Wallis (1972) proposa une statistique analogue à la statistique  $d$  pour tester un processus AR(4) simple. Sa **statistique  $d_4$**  est

$$d_4 = \frac{\sum_{t=5}^n (\tilde{u}_t - \tilde{u}_{t-4})^2}{\sum_{t=1}^n \tilde{u}_t^2}, \quad (10.84)$$

et ses propriétés sont très similaires à celles de la statistique  $d$  originelle. Lorsque le modèle est linéaire, qu'il n'y a pas de variable dépendante retardée, et que l'échantillon est petit, on peut utiliser  $d_4$  à la place du test standard, fondé sur la GNR, qui implique la régression de  $\tilde{\mathbf{u}}$  sur  $\tilde{\mathbf{X}}$  et sur  $\tilde{\mathbf{u}}_{-4}$ .

Un type très différent de test, qu'il appelle "optimal en un point", parce que ce type est conçu pour tester contre une alternative *simple*, fut proposé par King (1985a). Il est basé sur le ratio de la somme des résidus au carré pour une régression avec des valeurs fixes de  $\rho$ , disons  $\rho = 0.5$  ou  $\rho = 0.75$ , par la SSR d'une régression sans autocorrélation. Des valeurs critiques peuvent se calculer par des méthodes similaires à celles utilisées pour calculer la distribution exacte de la statistique  $d$ . Il est évident que ce test peut, comme son nom le suggère, avoir plus de puissance que les tests conventionnels lorsque la vraie valeur de  $\rho$  est à la fois assez différente de zéro et peu éloignée de la valeur de l'hypothèse utilisée dans le calcul de la statistique de test. King (1985b), King et McAleer (1987), Dastoor et Fisher (1988), et Dufour et King (1991) sont d'autres références que l'on peut citer sur les tests optimaux en un point.

Dans la littérature consacrée aux séries temporelles, de nombreux tests d'autocorrélation des résidus furent proposés. Deux d'entre eux largement utilisés sont les tests proposés par Box et Pierce (1970) et Ljung et Box (1978), qui sont tous deux fondés sur les **autocorrélations des résidus**, c'est-à-dire les corrélations entre  $\tilde{u}_t$  et  $\tilde{u}_{t-1}$ ,  $\tilde{u}_t$  et  $\tilde{u}_{t-2}$ , et ainsi de suite, jusqu'à atteindre des retards suffisamment éloignés. Ces tests sont valables lorsqu'ils sont utilisés dans leur objectif originel, qui reste le test des modèles ARIMA pour les tests d'autocorrélation des résidus, mais ils ne sont en général pas valables lorsqu'ils sont exécutés avec des résidus provenant de modèles de régression linéaire ou non linéaire qui contiennent à la fois des variables exogènes et des variables dépendantes retardées dans leurs fonctions de régression. La raison pour laquelle ils sont non valables dans ces circonstances est essentiellement la même raison qui fait que la statistique  $d$  n'est pas valable lorsqu'il y a des variables dépendantes retardées parmi les régresseurs; consulter Poskitt et Tremayne (1981).

## 10.9 CONTRAINTES DU FACTEUR COMMUN

Si la fonction de régression est mal spécifiée, les résidus peuvent manifester une autocorrélation même lorsque les aléas sont en réalité indépendants entre eux. Cela peut survenir si une variable qui était elle-même autocorrélée, ou une variable dépendante retardée, était omise de la fonction de régression. Dans un tel cas, on peut en général pratiquer des inférences valides en éliminant simplement l'erreur de spécification plutôt qu'en essayant de "corriger" le modèle en y intégrant des aléas AR(1) ou suivant tout autre processus. Si on effectuait la seconde manipulation, comme cela est trop fréquemment le cas dans les travaux appliqués, cela pourrait nous conduire à une grave erreur de spécification du modèle.

Il n'existe aucun moyen universel efficace, qui empêcherait la mauvaise interprétation d'une erreur de spécification de la fonction de régression telle que la présence d'aléas autocorrélés. La spécification d'un modèle tient plus de l'art que de la science, et avec des échantillons de séries temporelles typiquement très petits on ne peut jamais espérer détecter toutes les formes de mauvaise spécification. Néanmoins, il existe une famille de tests qui a donné des preuves d'efficacité dans la détection d'une mauvaise spécification dans les modèles où les aléas paraissent suivre un processus AR d'ordre peu élevé. Ce sont les tests que l'on appelle généralement, pour des raisons qui apparaîtront dans un instant, des tests des **contraintes du facteur commun**. L'idée de base de ces tests des contraintes du facteur commun se retrouve chez Sargan (1964). Les références les plus récentes sont Hendry et Mizon (1978), Mizon et Hendry (1980), et Sargan (1980a). Hendry (1980) fournit un exemple révélateur qui présente un modèle exagérément mal spécifié et qui entraîne en apparence des résultats cohérents après une "correction" par des aléas AR(1). Il montre ensuite qu'un test des contraintes du facteur commun détecterait l'erreur de spécification.

Dans le but de fixer les idées, nous allons supposer pour l'instant que le modèle qu'il faut tester est un modèle de régression linéaire doté en apparence d'erreurs AR(1). Il est naturel de penser à l'existence de *trois* modèles emboîtés dans ce cas. Le premier est le modèle de régression linéaire d'origine pour lequel les aléas sont supposés indépendants entre eux,

$$H_0 : y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \quad (10.85)$$

Le deuxième modèle est le modèle non linéaire qui résulte de la prise en compte d'aléas  $u_t$  de (10.85) suivant le processus AR(1)  $u_t = \rho u_{t-1} + \varepsilon_t$ ,

$$H_1 : y_t = \mathbf{X}_t\boldsymbol{\beta} + \rho(y_{t-1} - \mathbf{X}_{t-1}\boldsymbol{\beta}) + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (10.86)$$

Le troisième modèle est le modèle linéaire qui résulte de (10.86) après avoir relâché les contraintes non linéaires:

$$H_2 : y_t = \mathbf{X}_t\boldsymbol{\beta} + \rho y_{t-1} + \mathbf{X}_{t-1}\boldsymbol{\gamma} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (10.87)$$

où  $\beta$  et  $\gamma$  sont tous deux des vecteurs de dimension  $k$ . Nous avons déjà vu  $H_2$ , dans la Section 10.3, où il était utilisé pour obtenir une estimation initiale convergente de  $\rho$ .

A condition que tous ces modèles soient estimés sur un échantillon comparable (sans doute des observations allant de 2 à  $n$ , puisque  $H_1$  et  $H_2$  ne peuvent pas être estimés avec l'observation 1), le modèle initial,  $H_0$ , est un cas particulier du modèle qui contient des aléas AR(1),  $H_1$ , qui est à son tour un cas particulier du modèle linéaire non contraint,  $H_2$ . Des tests d'autocorrélation, tels que ceux dont nous avons discuté au cours de la Section 10.8, sont conçus pour tester  $H_0$  contre  $H_1$ . Si un tel test rejette l'hypothèse nulle, ce peut être le cas parce que  $H_1$  a en réalité généré les données, mais cela peut aussi être le cas parce que le modèle est mal spécifié. Le test de  $H_1$  contre  $H_2$  est un moyen de voir si le premier est un modèle acceptable. Ceci est un exemple du test des contraintes du facteur commun.

Il est naturel de se demander pourquoi les contraintes que (10.86) fait porter sur (10.87) sont appelées contraintes du *facteur commun*. Avec la notation avec l'opérateur retard, on peut écrire (10.86) comme

$$(1 - \rho L)y_t = (1 - \rho L)\mathbf{X}_t\beta + \varepsilon_t \quad (10.88)$$

et (10.78) peut s'écrire

$$(1 - \rho L)y_t = \mathbf{X}_t\beta + L\mathbf{X}_t\gamma + \varepsilon_t. \quad (10.89)$$

Il est évident que dans (10.88), mais pas dans (10.89), le facteur commun  $1 - \rho L$  apparaît dans les deux membres de l'équation. Cela explique le nom donné aux contraintes.

Bien que notre traitement se focalise sur le cas AR(1), les contraintes du facteur commun sont implicites dans les modèles de régression linéaire à aléas autorégressifs à un ordre quelconque. Par exemple, un modèle de régression linéaire à aléas AR(2) peut s'écrire

$$(1 - \rho_1 L - \rho_2 L^2)y_t = (1 - \rho_1 L - \rho_2 L^2)\mathbf{X}_t\beta + \varepsilon_t, \quad (10.90)$$

alors que la variante non contrainte correspondant à (10.89) peut s'écrire

$$(1 - \rho_1 L - \rho_2 L^2)y_t = \mathbf{X}_t\beta + L\mathbf{X}_t\gamma_1 + L^2\mathbf{X}_t\gamma_2 + \varepsilon_t, \quad (10.91)$$

où  $\gamma_1$  et  $\gamma_2$  sont des vecteurs à  $k$  composantes. A nouveau, on voit que le facteur  $1 - \rho_1 L - \rho_2 L^2$  apparaît dans les deux membres de l'équation (10.90) mais uniquement dans le membre de gauche de (10.91). Les tests des contraintes du facteur commun dans les modèles à processus AR d'ordre supérieur sont pour l'essentiel les mêmes que les tests pour les modèles à erreurs AR(1); pour être simple, notre traitement de ces tests ne considérera que le cas AR(1).

Dans la plupart des cas, le moyen le plus simple de tester des contraintes du facteur commun, et probablement aussi le moyen le plus fiable, consiste à utiliser un test en  $F$  asymptotique. Ainsi la statistique qui permet de tester  $H_1$  contre  $H_2$ , c'est-à-dire (10.86) contre (10.87), serait

$$\frac{(\text{SSR}_1 - \text{SSR}_2)/l}{\text{SSR}_2/(n - k - l - 2)}, \quad (10.92)$$

où  $\text{SSR}_1$  est la somme des résidus au carré de l'estimation par moindres carrés de  $H_1$ ,  $\text{SSR}_2$  est la somme des résidus au carré de l'estimation par moindres carrés de  $H_2$ , et où  $l \leq k$  est le nombre de degrés de liberté pour le test. Les degrés de liberté du dénominateur sont au nombre de  $n - k - l - 2$  parce que  $H_2$  est estimé sur  $n - 1$  observations, et possède  $k + 1 + l$  paramètres, correspondant aux  $k$   $\beta_i$ , à  $\rho$ , et aux  $l$  paramètres additionnels. Remarquons que bien que ce test soit parfaitement valable asymptotiquement, il ne sera pas exact avec des échantillons finis, sans considération quelconque de la distribution des  $\varepsilon_t$ , parce qu'aussi bien  $H_1$  que  $H_2$  comprennent des variables dépendantes retardées dans le membre de droite et également parce que  $H_1$  est non linéaire en ses paramètres.

Nous en venons désormais à un aspect des tests du facteur commun légèrement compliqué: la détermination du nombre de contraintes,  $l$ . Dans le cas précédent du test de  $H_1$  contre  $H_2$ , il *semble* qu'il y ait  $k$  contraintes. Après tout,  $H_1$  possède  $k + 1$  paramètres (les  $k$   $\beta_i$  et  $\rho$ ) et  $H_2$  semble posséder  $2k + 1$  paramètres (les  $k$   $\beta_i$ , les  $k$   $\gamma_i$ , et  $\rho$ ). La différence est  $(2k + 1) - (k + 1)$ , c'est-à-dire  $k$ . En réalité cependant, le nombre de contraintes sera presque toujours *inférieur* à  $k$ , parce que, à l'exception de rares cas, le nombre de paramètres *identifiables* dans  $H_2$  sera inférieur à  $2k + 1$ . Et le meilleur moyen de voir pourquoi cela sera toujours le cas est de considérer un exemple.

Supposons que la fonction de régression  $x_t(\beta)$  du modèle d'origine  $H_0$  soit

$$\beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 z_{t-1} + \beta_4 y_{t-1}, \quad (10.93)$$

où  $z_t$  est la  $t^{\text{ième}}$  observation d'une variable économique temporelle, et  $t$  est la  $t^{\text{ième}}$  observation d'une tendance temporelle linéaire. La fonction de régression pour le modèle non contraint  $H_2$  qui correspond à (10.93) est

$$\begin{aligned} &\beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 z_{t-1} + \beta_4 y_{t-1} + \rho y_{t-1} \\ &+ \gamma_0 + \gamma_1 z_{t-1} + \gamma_2 (t - 1) + \gamma_3 z_{t-2} + \gamma_4 y_{t-2}. \end{aligned} \quad (10.94)$$

Cette fonction de régression s'avère posséder 11 paramètres, mais en réalité 4 d'entre eux ne sont pas identifiables. Il est évident que l'on ne peut pas estimer à la fois  $\beta_0$  et  $\gamma_0$ , car il ne peut y avoir qu'un unique terme constant. De manière similaire, on ne peut pas estimer simultanément  $\beta_3$  et  $\gamma_1$ , puisqu'il ne peut pas y avoir deux coefficients associés à  $z_{t-1}$ , et on ne peut pas estimer à la fois  $\beta_4$  et  $\rho$ , puisqu'il ne peut pas y avoir deux coefficients associés à  $y_{t-1}$ .

Enfin on ne peut pas estimer  $\gamma_2$  en même temps que  $\beta_2$  et que la constante parce que  $t$ ,  $t - 1$  et le terme constant sont parfaitement colinéaires, puisque  $t - (t - 1) = 1$ . Ainsi, la version de  $H_2$  que l'on peut en fait estimer est caractérisée par la fonction de régression

$$\delta_0 + \beta_1 z_t + \delta_1 t + \delta_2 z_{t-1} + \delta_3 y_{t-1} + \gamma_3 z_{t-2} + \gamma_4 y_{t-2}, \quad (10.95)$$

où

$$\delta_0 = \beta_0 + \gamma_0 - \gamma_2; \quad \delta_1 = \beta_2 + \gamma_2; \quad \delta_2 = \beta_3 + \gamma_1; \quad \text{et} \quad \delta_3 = \rho + \beta_4.$$

Nous voyons que (10.95) possède 7 paramètres que l'on peut estimer:  $\beta_1$ ,  $\gamma_3$ ,  $\gamma_4$ , et de  $\delta_0$  à  $\delta_3$ , au lieu des 11 paramètres de (10.94), dont beaucoup d'entre eux ne sont pas identifiables. La fonction de régression pour le modèle contraint  $H_1$  est

$$\begin{aligned} & \beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 z_{t-1} + \beta_4 y_{t-1} + \rho y_{t-1} \\ & - \rho \beta_0 - \rho \beta_1 z_{t-1} - \rho \beta_2 (t - 1) - \rho \beta_3 z_{t-2} - \rho \beta_4 y_{t-2}, \end{aligned}$$

et elle possède six paramètres,  $\rho$  et de  $\beta_0$  à  $\beta_4$ . Ainsi dans ce cas,  $l$ , le nombre de contraintes imposées à  $H_2$  par  $H_1$  est égal à 1.

Bien que ce soit un exemple assez extrême, des problèmes comparables surviennent dans presque toute tentative de test des contraintes du facteur commun. Les termes constants, de nombreux types de variables muettes (en particulier les variables muettes saisonnières et des tendances temporelles), des variables dépendantes retardées, et des variables indépendantes qui apparaissent avec plus d'un indice chronologique se retrouvent presque inmanquablement dans un modèle non contraint  $H_2$  pour lequel tous les paramètres ne sont pas identifiables. Par chance, il est très facile de traiter ce genre de problème lorsque l'on utilise un test en  $F$ ; il suffit simplement d'omettre les régresseurs redondants au moment d'estimer  $H_2$ . On peut ensuite calculer  $l$  comme la différence entre le nombre de paramètres de  $H_2$  et celui de  $H_1$ , à savoir  $k + 1$ . Puisque de nombreux progiciels de régression ignorent automatiquement les régresseurs redondants, une approche naïve mais souvent efficace consiste simplement à essayer d'estimer  $H_2$  sous une forme proche du modèle originel et de compter ensuite le nombre de paramètres que le progiciel a été capable d'estimer.

Le test en  $F$  (10.92) n'est pas l'unique moyen de tester les contraintes du facteur commun. Puisque la fonction de régression pour  $H_2$  est linéaire en ses paramètres, alors que celle de  $H_1$  est non linéaire, il est naturel d'essayer de baser les tests sur les estimations OLS de  $H_2$  uniquement. Sargan (1980a) discute d'une approche de ce problème mais elle est assez difficile, et nécessite un logiciel spécifique. Une approche plus simple consiste à utiliser un estimateur en une étape de  $H_1$ . On peut obtenir des estimations convergentes des paramètres de  $H_1$  à partir des estimations de  $H_2$ , ainsi que nous en avons

parlé dans la Section 10.3, et la GNR est ensuite utilisée pour aboutir aux estimations en une étape. Ces paramètres en eux-mêmes ne sont pas intéressants. Tout ce dont nous avons besoin, c'est de la SSR de la GNR, que l'on peut utiliser à la place de  $SSR_1$  dans la formule pour le test en  $F$ . Cependant, comme il n'est en général ni difficile ni trop coûteux d'estimer  $H_1$  à l'aide des ordinateurs et des progiciels modernes, les circonstances pour lesquelles un avantage conséquent en faveur de la procédure en une étape apparaît sont probablement rares.

On peut utiliser une procédure comparable à un test des contraintes du facteur commun lorsque le modèle originel ( $H_0$ ) est non linéaire. Dans ce contexte, le modèle  $H_1$  peut s'écrire

$$(1 - \rho L)y_t = (1 - \rho L)x_t(\beta) + \varepsilon_t. \quad (10.96)$$

Une version de (10.96) pour laquelle la contrainte du facteur commun n'est pas valable est

$$(1 - \rho L)y_t = (1 - \delta L)x_t(\beta) + \varepsilon_t. \quad (10.97)$$

Evidemment, (10.96) est simplement (10.97) soumis à la contrainte d'égalité entre  $\delta$  et  $\rho$ . La régression de Gauss-Newton permet de tester cette contrainte de la manière habituelle. Cette GNR est

$$\begin{aligned} \mathbf{y} - \hat{\mathbf{x}} - \hat{\rho}(\mathbf{y}_{-1} - \hat{\mathbf{x}}_{-1}) &= (\hat{\mathbf{X}} - \hat{\rho}\hat{\mathbf{X}}_{-1})\mathbf{b} \\ &+ r(\mathbf{y}_{-1} - \hat{\mathbf{x}}_{-1}) + d\hat{\mathbf{x}}_{-1} + \text{résidus}, \end{aligned} \quad (10.98)$$

où  $\hat{\rho}$  et  $\hat{\beta}$  sont les paramètres de  $H_1$  estimés par NLS, et où  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\beta})$ . La régression (10.98) ressemble exactement à la GNR (10.26), dont nous avons fait usage pour calculer la matrice de covariance de  $\hat{\beta}$  et  $\hat{\rho}$ , à laquelle on aurait ajouté le régresseur  $\hat{\mathbf{x}}_{-1}$ , dont le coefficient est  $d$ . Le  $t$  de Student pour  $d = 0$  sera une statistique de test asymptotiquement valable.

Notons que cette GNR pourrait être employée même si  $x_t(\beta)$  était une fonction linéaire. Puisque cette variante du test des contraintes du facteur commun ne possède qu'un unique degré de liberté, elle serait différente de la forme habituelle du test pour tout modèle où  $l > 1$ . La différence apparaît parce que le test basé sur (10.98) teste contre un modèle alternatif moins général que ne le fait le test sous sa forme habituelle. Lorsque  $x_t(\beta)$  est linéaire, on peut écrire (10.97) comme

$$(1 - \rho L)y_t = \mathbf{X}_t\beta - \delta\mathbf{X}_{t-1}\beta + \varepsilon_t, \quad (10.99)$$

qui est en général plus restrictif que l'équation (10.89) (sauf lorsque  $l = 1$ ). Ainsi la prise en compte du cas de la régression non linéaire révèle qu'il y a en réalité deux tests des contraintes du facteur commun différents lorsque le modèle est linéaire. Le premier test, qui teste (10.88) contre (10.89), est le test en  $F$  (10.92). Il aura  $l$  degrés de liberté, où  $1 \leq l \leq k$ . Le second, qui



teste (10.88) contre (10.99), est le test en  $t$  pour  $d = 0$  dans la régression de Gauss-Newton (10.98). Il possédera toujours un seul degré de liberté. Selon la manière dont les données ont été générées, chacun des deux tests peut être meilleur que l'autre; consulter le Chapitre 12. Lorsque  $l = 1$ , les deux tests coïncident, et cela serait un bon exercice de le démontrer.

## 10.10 VARIABLES INSTRUMENTALES ET AUTOCORRÉLATION

Jusqu'à présent au cours de ce chapitre, nous avons supposé que la fonction de régression  $\mathbf{x}(\boldsymbol{\beta})$  dépendait uniquement de variables exogènes et prédéterminées. Cependant, il n'y a pas de raison pour que des aléas autocorrélés n'apparaissent pas dans des modèles pour lesquels des variables endogènes courantes apparaissent dans la fonction de régression. Comme nous l'avons vu dans le Chapitre 7, la technique d'estimation par variables instrumentales (IV) est fréquemment employée pour obtenir des estimations convergentes pour de tels modèles. Dans cette section, nous discuterons brièvement de la façon d'utiliser les méthodes IV pour estimer des modèles de régression univariée dont les aléas sont autocorrélés dans de tels modèles.

Supposons que l'on veuille estimer le modèle (10.12) à l'aide des variables instrumentales. Alors, comme nous l'avons vu dans la Section 7.6, les estimations IV peuvent s'obtenir en minimisant, par rapport à  $\boldsymbol{\beta}$ , la fonction critère

$$(\mathbf{y} - \mathbf{x}'(\boldsymbol{\beta}))^\top \mathbf{P}_W (\mathbf{y} - \mathbf{x}'(\boldsymbol{\beta})), \quad (10.100)$$

où la fonction de régression  $\mathbf{x}'(\boldsymbol{\beta})$  est définie par (10.13), et  $\mathbf{P}_W$  est la matrice qui projette orthogonalement sur l'espace engendré par les colonnes de  $\mathbf{W}$ , une matrice d'instruments adéquate. La forme IV de la régression de Gauss-Newton peut servir de base à un algorithme de minimisation de (10.100). Etant données les conditions de régularité convenables pour  $x_t(\boldsymbol{\beta})$ , et en supposant que  $|\rho| < 1$ , ces estimations seront convergentes et asymptotiquement normales. Consulter Sargan (1959) pour un traitement complet du cas pour lequel  $\mathbf{x}(\boldsymbol{\beta})$  est linéaire.

La seule difficulté éventuelle rencontrée avec cette procédure IV est qu'il faut trouver une matrice d'instruments  $\mathbf{W}$  "adéquate". Pour garantir l'efficacité asymptotique, on veut toujours que les instruments contiennent toutes les variables exogènes et prédéterminées qui apparaissent dans la fonction de régression. A partir de (10.13), nous voyons que davantage de ces variables apparaissent dans la fonction de régression  $x'_t(\boldsymbol{\beta})$  pour le modèle transformé que dans la fonction de régression d'origine  $x_t(\boldsymbol{\beta})$ . Ainsi le choix optimal des instruments peut différer selon que l'on prend ou non en compte l'autocorrélation.

Pour rendre cet aspect plus clair, supposons que le modèle originel soit linéaire, avec la fonction de régression

$$x_t(\boldsymbol{\beta}) = \mathbf{Z}_t \boldsymbol{\beta}_1 + \mathbf{Y}_t \boldsymbol{\beta}_2, \quad (10.101)$$

où  $\mathbf{Z}_t$  est un vecteur composé de variables explicatives exogènes ou prédéterminées, et  $\mathbf{Y}_t$  est un vecteur ligne composé de variables endogènes courantes; la dimension du vecteur  $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}_1 : \boldsymbol{\beta}_2]$  est  $k$ . La fonction de régression pour le modèle transformé est donc

$$x'_t(\boldsymbol{\beta}) = \rho y_{t-1} + \mathbf{Z}_t \boldsymbol{\beta}_1 + \mathbf{Y}_t \boldsymbol{\beta}_2 - \rho \mathbf{Z}_{t-1} \boldsymbol{\beta}_1 - \rho \mathbf{Y}_{t-1} \boldsymbol{\beta}_2. \quad (10.102)$$

Dans (10.101), les seules variables exogènes ou prédéterminées étaient les variables regroupées dans  $\mathbf{Z}_t$ . Toutefois, dans (10.102), elles correspondent à  $y_{t-1}$  et aux variables regroupées dans  $\mathbf{Z}_t$ ,  $\mathbf{Z}_{t-1}$ , et  $\mathbf{Y}_{t-1}$  (les mêmes variables peuvent apparaître dans plusieurs de ces éléments; voir la dimension sur les contraintes du facteur commun de la section précédente). Toutes ces variables seraient normalement comprises dans la matrice des instruments  $\mathbf{W}$ . Puisque ces variables sont en nombre presque certainement supérieur à  $k + 1$ , il ne serait pas nécessaire normalement d'ajouter des instruments pour garantir que tous les paramètres seront identifiés.

Pour une discussion plus approfondie sur l'estimation d'une seule équation linéaire à aléas autocorrélés et avec régresseurs endogènes courants, consulter Sargan (1959, 1961), Amemiya (1966), Fair (1970), Dhrymes, Berner, et Cummins (1974), Hatanaka (1976), et Bowden et Turkington (1984).

Les tests d'autocorrélation dans les modèles estimés par IV sont immédiats si l'on emploie une variante de la régression de Gauss-Newton. Dans la Section 7.7, nous avons discuté de la GNR (7.37), pour laquelle la régressande et les régresseurs sont évalués avec les estimations contraintes, et nous avons montré comment l'utiliser pour calculer des statistiques de test. Les tests d'autocorrélation sont simplement des applications de cette procédure. Supposons que l'on veuille tester un modèle rendu non linéaire par des aléas AR(1). Le modèle alternatif est donné par (10.12), pour les observations allant de 2 à  $n$ , avec une hypothèse nulle correspondant à  $\rho = 0$ . Dans ce cas, la GNR (7.38) est

$$\tilde{\mathbf{u}} = \mathbf{P}_W \tilde{\mathbf{X}} \mathbf{b} + r \mathbf{P}_W \tilde{\mathbf{u}}_{-1} + \text{résidus}, \quad (10.103)$$

où  $\tilde{\boldsymbol{\beta}}$  désigne les estimations IV sous l'hypothèse nulle d'absence d'autocorrélation,  $\tilde{\mathbf{u}}$  désigne  $\mathbf{y} - \mathbf{x}(\tilde{\boldsymbol{\beta}})$ , et  $\tilde{\mathbf{X}}$  désigne  $\mathbf{X}(\tilde{\boldsymbol{\beta}})$ . C'est à l'évidence l'analogue IV de la régression (10.76); si les deux manifestations de  $\mathbf{P}_W$  étaient annulées, (10.76) et (10.103) seraient identiques. Le  $t$  de Student de l'estimation de  $r$  à partir de cette régression sera une statistique de test valable. Cela sera exact aussi bien lorsque (10.103) est estimé de façon explicite par OLS que lorsque  $\tilde{\mathbf{u}}$  est régressé sur  $\tilde{\mathbf{X}}$  et sur  $\tilde{\mathbf{u}}_{-1}$  à l'aide d'une procédure IV avec  $\mathbf{W}$  en tant que matrice d'instruments. Malgré tout, lorsqu'une régression artificielle comparable à (10.103) est utilisée pour tester des autocorrélations à un ordre plus élevé, la régression doit être estimée de façon explicite par OLS si l'on veut produire un test en  $F$  valable. Nous avons déjà discuté de tout cela lors de la Section 7.7.

Comme d'habitude, il y a deux résultats mineurs à établir avant que cette procédure ne soit mise en œuvre. Premièrement, il y a le problème de la gestion de la première observation. L'approche la plus simple consiste sans doute à la conserver et à établir le résidu non observé  $\tilde{u}_0$  à zéro dans l'intention d'utiliser la GNR (10.103), mais il existe d'autres éventualités qui produiront des résultats différents avec des échantillons finis; consulter la Section 10.8.

En second lieu, il y a le problème du choix des instruments à utiliser lors de l'exécution de la GNR (10.103). Si nous voulions minimiser la fonction critère (10.100) pour obtenir des estimations simultanées de  $\beta$  et  $\rho$ , alors, comme nous l'avons vu, il serait généralement préférable d'employer davantage d'instruments que pour obtenir  $\tilde{\beta}$ . Identiquement, il serait souhaitable lors du test de l'hypothèse  $\rho = 0$  que  $\mathbf{W}$  comprenne à la fois  $\mathbf{y}_{-1}$  et les régresseurs qui apparaissent dans  $\mathbf{x}_{-1}(\tilde{\beta})$ . Dans ce cas, comme nous l'avons vu dans la Section 7.7, la statistique de test doit se calculer comme un test pseudo- $t$  ou pseudo- $F$  basé sur le principe  $C(\alpha)$ .

Pour plus de détails sur les tests d'autocorrélation dans les modèles estimés par IV, consulter Godfrey (1976, 1988), Harvey et Phillips (1980, 1981), et Sargan et Mehta (1983).

## 10.11 AUTOCORRÉLATION ET MODÈLES MULTIVARIÉS

Nous avons vu des modèles de régression multivariée dans la Section 9.7. Lorsqu'on estime ces modèles à l'aide de données chronologiques il faut s'attendre à ce qu'ils manifestent de l'autocorrélation. Les méthodes relatives à l'estimation et aux tests des modèles multivariés avec autocorrélation sont pour une bonne part des combinaisons évidentes des techniques déjà rencontrées dans ce chapitre et de celles discutées dans la Section 9.7. Il y a, cependant, un petit nombre d'aspects nouveaux à ce problème, et nous les traiterons dans cette section.

Considérons la classe de modèles

$$\mathbf{y}_t = \xi_t(\beta) + \mathbf{u}_t, \quad \mathbf{u}_t = \mathbf{u}_{t-1}\mathbf{R} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(\mathbf{0}, \mathbf{\Omega}), \quad (10.104)$$

où  $\mathbf{y}_t$ ,  $\xi_t(\beta)$ ,  $\mathbf{u}_t$ , et  $\varepsilon_t$  sont des vecteurs à  $1 \times m$  composantes, et où  $\mathbf{R}$  et  $\mathbf{\Omega}$  sont des matrices de dimension  $m \times m$ . Ceci définit la famille générale des modèles de régression multivariée à aléas AR(1). Il est conceptuellement très aisé de transformer (10.104) en

$$\mathbf{y}_t = \xi_t(\beta) + \mathbf{y}_{t-1}\mathbf{R} - \xi_{t-1}(\beta)\mathbf{R} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(\mathbf{0}, \mathbf{\Omega}), \quad (10.105)$$

et nous traiterons désormais (10.105) comme n'importe quel autre modèle de régression multivariée. Mais remarquons qu'au lieu d'un scalaire  $\rho$  nous avons maintenant une matrice  $\mathbf{R}$  de dimension  $m \times m$ , ce qui permet à chaque

élément de  $\mathbf{u}_t$  de dépendre de chaque élément de  $\mathbf{u}_{t-1}$ . Ainsi, si  $m$  est important, même la possibilité d'une autocorrélation réduite à l'ordre un introduit à l'évidence un nombre important de paramètres additionnels, ce qui peut être nuisible à l'obtention d'estimations fiables des paramètres  $\beta$  et  $\mathbf{R}$ . Dans le but de réduire ce nombre de paramètres à estimer, certains voudront imposer des contraintes sur  $\mathbf{R}$ . Une contrainte naturelle est que ce soit une matrice diagonale, ce qui implique que  $u_{ti}$  dépende de  $u_{t-1,i}$  et non pas de  $u_{t-1,j}$  pour tout  $j \neq i$ .

Au cours de la Section 9.7, nous avons discuté et donné des exemples de systèmes d'équations singuliers, pour lesquels les aléas sont contraints à avoir une somme nulle à travers toutes les équations. Nous avons vu que de tels systèmes apparaissent très fréquemment dans la pratique. Berndt et Savin (1975) ont démontré que si un système d'équations est singulier, cela revenait à assigner des contraintes strictes sur la forme que  $\mathbf{R}$  peut prendre. En particulier, si  $\mathbf{R}$  est supposée diagonale, tous les éléments doivent alors être égaux. Pour en comprendre la raison, supposons simplement que  $m = 2$  et écrivons le processus AR  $\mathbf{u}_t = \mathbf{u}_{t-1}\mathbf{R} + \varepsilon_t$  comme

$$\begin{aligned} u_{t1} &= r_{11}u_{t-1,1} + \varepsilon_{t1} \\ u_{t2} &= r_{22}u_{t-1,2} + \varepsilon_{t2}. \end{aligned}$$

En sommant ces équations, on voit que

$$u_{t1} + u_{t2} = r_{11}u_{t-1,1} + r_{22}u_{t-1,2} + \varepsilon_{t1} + \varepsilon_{t2}. \quad (10.106)$$

Par hypothèse,  $u_{t-1,1} + u_{t-1,2} = 0$  et  $\varepsilon_{t1} + \varepsilon_{t2} = 0$ . Mais ces deux contraintes impliquent que  $u_{t1} + u_{t2} = 0$  uniquement si  $r_{11} = r_{22} = \rho$ . Si c'est le cas, on peut écrire (10.106) comme

$$u_{t1} + u_{t2} = \rho(u_{t-1,1} + u_{t-1,2}) + \varepsilon_{t1} + \varepsilon_{t2} = \rho \cdot 0 + 0 = 0.$$

Ainsi, lorsque  $r_{11} = r_{22} = \rho$ , il est aisé de voir que si la somme des  $\varepsilon_{ti}$  est nulle, celle des  $u_{ti}$  le sera aussi; imaginons que l'on débute avec  $u_{0i} = 0$  puis que l'on résolve de manière récursive.

Le résultat de Berndt-Savin, qui se généralise bien sûr à des matrices  $\mathbf{R}$  non diagonales et à des processus AR de plus haut degré, signifie qu'il faut être prudent lors de la spécification des processus temporels pour prendre en considération les aléas des systèmes d'équations singuliers. Si l'on spécifie malencontreusement une matrice  $\mathbf{R}$  qui ne satisfait pas les contraintes de Berndt-Savin, le système transformé (10.105) ne sera plus singulier, et le résultat est que l'on obtiendra des estimations des paramètres différentes en mettant à l'écart des équations différentes. D'autre part, le fait que si  $\mathbf{R}$  est diagonale tous les éléments diagonaux sont identiques nous autorise des simplifications considérables dans certains cas. Beach et MacKinnon (1979) utilisent ce

résultat pour développer un estimateur ML qui conserve la première observation pour des systèmes d'équations singuliers avec aléas AR(1) et matrice  $\mathbf{R}$  diagonale.

Cette section fut plutôt brève. On trouvera une discussion plus détaillée sur l'autocorrélation dans les modèles multivariés ainsi que de nombreuses références dans l'ouvrage de Srivastava et Giles (1987, Chapitre 7).

## 10.12 CONCLUSION

Malgré la longueur de ce chapitre, nous n'avons aucunement couvert tout le domaine de l'autocorrélation. Notre discussion sur les résultats des séries chronologiques fut délibérément concise: les lecteurs qui ne sont pas à l'aise avec cette littérature voudront sûrement consulter l'ouvrage de Harvey (1981, 1989), Granger et Newbold (1986), ou un des nombreux ouvrages plus pointus cités par Granger et Watson (1984). Un certain nombre de thèmes étroitement rattachés à ceux traités dans ce chapitre sera développé dans les Chapitres 19 et 20.

Dans ce chapitre, nous avons essayé de mettre l'accent sur les tests de spécification, et principalement sur ceux basés sur la régression de Gauss-Newton. Un certain nombre d'autres tests de spécification basés sur la GNR, dont certains peuvent être interprétés comme des alternatives à des tests d'autocorrélation, dont la plupart sont adaptables à des modèles qui incorporent une transformation pour tenir compte de l'autocorrélation, sera exposé dans le Chapitre 11. La manière d'interpréter les résultats des tests de spécification tels que ceux-ci fera l'objet du Chapitre 12. Tous les tests d'autocorrélation et des contraintes du facteur commun présentés dans ce chapitre deviennent plus compréhensibles dans le contexte des résultats qui y seront établis.

## TERMES ET CONCEPTS

alternative localement équivalente	procédure alternative de Durbin
autocorrélation	procédure de Cochrane-Orcutt
autocorrélation de résidus	procédure de Hildreth-Lu
condition d'inversibilité	processus à aléas autorégressifs
condition de stationnarité	processus d'aléas moyenne mobile
contraintes du facteur commun	processus AR(1), AR(2), AR(4), et AR( $p$ )
estimation ML complète	processus AR(4) simple
estimations en une étape	processus ARIMA( $p, d, q$ )
grille de recherche	processus ARMA( $p, q$ )
indépendance (des aléas)	processus AR saisonnier
innovation	processus MA(1) et MA( $q$ )
méthodes pour séries chronologiques	
opérateur retard	

processus moyenne mobile  
  autorégressif  
racines à l'extérieur du cercle de  
  rayon 1  
recherche en alternance  
région de stationnarité

séries temporelles intégrées  
statistique  $d$  (statistique DW )  
statistique  $d_4$   
test en  $h$  de Durbin  
triangle de stationnarité  
transformation de Prais-Winsten

# Chapitre 11

## Tests Basés sur la Régression de Gauss-Newton

### 11.1 INTRODUCTION

Dans la Section 6.4, nous avons montré que la régression de Gauss-Newton offrait un moyen simple de tester des contraintes sur les paramètres d'une fonction de régression dès que l'on disposait des estimations convergentes au taux  $n^{1/2}$  de ces paramètres qui satisfont les contraintes. Dans la plupart des cas, elles correspondent aux estimations par moindres carrés du modèle contraint. Dans la Section 10.8, nous avons montré que l'on pouvait exécuter les tests pour à peu près tous les genres de corrélations en série grâce à des variantes de la GNR. Au cours de ce chapitre, nous discuterons de nombreux tests complémentaires basés sur la GNR qui peuvent se révéler d'une grande utilité dans les études économétriques appliquées. Ces tests sont:

- (i) des tests d'égalité de deux (ou plus) ensembles de paramètres;
- (ii) des tests d'hypothèses de modèles non emboîtés, pour lesquels un modèle de régression est testé contre un ou plusieurs modèles alternatifs non emboîtés;
- (iii) des tests basés sur la comparaison de deux ensembles d'estimations, dont l'un est généralement convergent sous des conditions moins fortes que l'autre;
- (iv) des tests d'hétéroscédasticité dont la forme est connue.

Dans la dernière section du chapitre, nous aborderons un matériau très important et qui sera traité en détail dans le Chapitre 16. La régression de Gauss-Newton n'est valable que sous l'hypothèse d'homoscédasticité des aléas, une hypothèse qui est quelquefois trop forte. Dans cette dernière section, nous discuterons d'une régression artificielle qui peut être utilisée pour calculer des statistiques de test à chaque fois que l'on peut utiliser la GNR, mais qui a la propriété avantageuse de fournir des statistiques de test asymptotiquement valables même lorsque les aléas manifestent un phénomène d'hétéroscédasticité dont la forme est inconnue. Nous présentons cette régression artificielle parce qu'il s'agit d'un prolongement logique de la

régression de Gauss-Newton, et parce qu'elle peut être très utile dans la pratique.

## 11.2 TESTS D'ÉGALITÉ DE DEUX VECTEURS DE PARAMÈTRES

L'un des problèmes classiques en économétrie consiste à savoir si les coefficients d'un modèle de régression (le plus souvent un modèle linéaire) sont identiques si l'on prend deux (ou quelquefois davantage) sous-échantillons distincts. Dans le cadre des séries temporelles, les sous-échantillons correspondraient généralement à des périodes différentes, et ces tests sont souvent appelés tests de **changement de régime**. Parfois nous désirons savoir si les coefficients sont identiques au cours de deux ou de plusieurs périodes dans le but de tester la bonne spécification du modèle. Dans de telles circonstances, les ensembles de données temporelles peuvent être divisés en deux périodes, la période actuelle et la période passée, de façon assez arbitraire pour les besoins du test. C'est une attitude légitime, mais de tels tests sont beaucoup plus intéressants lorsqu'il existe une raison de croire que les sous-échantillons correspondent à des conjonctures économiques bien distinctes, telles que les modifications de taux de change ou de régimes politiques.<sup>1</sup> Dans le cadre des données en coupe transversale, une division arbitraire n'est presque jamais pertinente; au lieu de cela, les sous-échantillons représenteraient des groupes potentiellement différents tels que les multinationales et les PME, les pays développés et les pays du tiers-monde, ou encore les hommes et les femmes. Dans ces cas, les résultats du test sont souvent intéressants en eux-mêmes. Par exemple, un économiste spécialisé dans le marché du travail peut être intéressé par les fonctions déterminant le salaire pour tester si ce sont les mêmes pour les hommes et pour les femmes, ou pour deux groupes ethniques différents.<sup>2</sup>

Un traitement traditionnel de ce problème prend ses sources dans la littérature statistique consacrée à l'analyse de la variance (Scheffé, 1959). En économétrie, c'est à G. C. Chow (1960) que l'on doit un article novateur et très influent, et par la suite le test en  $F$  habituel pour l'égalité de deux ensembles de coefficients dans les modèles de régression linéaire est souvent appelé le **test de Chow**. Fisher (1970) fournit un exposé plus clair de la procédure du test de Chow classique. Dufour (1982) fournit un exposé plus géométrique et

<sup>1</sup> Lorsqu'il n'y a pas de raison de croire à une modification des paramètres à une date quelconque, il peut être pertinent d'utiliser une procédure qui ne fait référence à aucune date. On utilisera par exemple les procédures CUSUM et CUSUM des carrés, de Brown, Durbin, et Evans (1975).

<sup>2</sup> Une fonction déterminant le salaire établit un lien entre les salaires et une série de variables explicatives telles que l'âge, la formation, et l'expérience. Pour des exemples d'utilisation de tests en  $F$  pour l'égalité de deux ensembles de coefficients dans ce contexte, consulter Oaxaca (1973, 1974).



fait une généralisation du test pour manipuler n'importe quel nombre de sous-échantillons, dont certains peuvent avoir un nombre d'observations inférieur au nombre de régresseurs.

La manière habituelle de poser le problème consiste à partitionner les données en deux ensembles, c'est-à-dire à partitionner le vecteur  $\mathbf{y}$  à  $n$  composantes de la variable dépendante en deux vecteurs  $\mathbf{y}_1$  et  $\mathbf{y}_2$ , respectivement de dimensions  $n_1$  et  $n_2$ , et à partitionner la matrice  $\mathbf{X}$  des observations sur les régresseurs de dimension  $n \times k$  en deux matrices  $\mathbf{X}_1$  et  $\mathbf{X}_2$ , qui sont respectivement de dimensions  $n_1 \times k$  et  $n_2 \times k$ . Cette partition nécessitera bien évidemment que les données soient ordonnées. Ainsi l'hypothèse maintenue peut s'écrire comme

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}, \quad (11.01)$$

où  $\boldsymbol{\beta}_1$  et  $\boldsymbol{\beta}_2$  sont des vecteurs à  $k$  paramètres qu'il faut estimer. L'hypothèse nulle que l'on teste est  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$ . Sous cette hypothèse nulle, l'équation (11.01) se réduit à

$$\mathbf{y} \equiv \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \equiv \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}. \quad (11.02)$$

Lorsqu'à la fois les tailles  $n_1$  et  $n_2$  sont supérieures au nombre de paramètres  $k$ , ce qui est le cas le plus courant, il est aisé de tester (11.01) contre (11.02) en faisant usage d'un test en  $F$  ordinaire tel que celui dont nous avons discuté à la Section 3.5. La somme des résidus au carré non contrainte qui résulte de l'estimation de (11.01) est

$$\text{USSR} = \mathbf{y}_1^\top \mathbf{M}_1 \mathbf{y}_1 + \mathbf{y}_2^\top \mathbf{M}_2 \mathbf{y}_2 = \text{SSR}_1 + \text{SSR}_2,$$

où  $\mathbf{M}_i \equiv \mathbf{I} - \mathbf{X}_i(\mathbf{X}_i^\top \mathbf{X}_i)^{-1} \mathbf{X}_i^\top$  pour  $i = 1, 2$ . Ainsi USSR correspond simplement à la somme de deux SSR correspondant respectivement aux régressions de  $\mathbf{y}_1$  sur  $\mathbf{X}_1$  et de  $\mathbf{y}_2$  sur  $\mathbf{X}_2$ . La SSR contrainte qui découle de l'estimation de (11.02) est

$$\text{RSSR} = \mathbf{y}^\top \mathbf{M}_X \mathbf{y},$$

où  $\mathbf{M}_X \equiv \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ . Ainsi la statistique  $F$  ordinaire est

$$\frac{(\mathbf{y}^\top \mathbf{M}_X \mathbf{y} - \mathbf{y}_1^\top \mathbf{M}_1 \mathbf{y}_1 - \mathbf{y}_2^\top \mathbf{M}_2 \mathbf{y}_2)/k}{(\mathbf{y}_1^\top \mathbf{M}_1 \mathbf{y}_1 + \mathbf{y}_2^\top \mathbf{M}_2 \mathbf{y}_2)/(n - 2k)} = \frac{(\text{RSSR} - \text{SSR}_1 - \text{SSR}_2)/k}{(\text{SSR}_1 + \text{SSR}_2)/(n - 2k)}. \quad (11.03)$$

Ce test comporte  $k$  et  $n - 2k$  degrés de liberté. Il y a  $k$  contraintes parce que le modèle contraint a  $k$  paramètres alors que le modèle non contraint en possède  $2k$ .

La statistique de test (11.03) est ce que de nombreux praticiens en économétrie croient être le test de Chow. Trois limites immédiates à ce test

se présentent. La première limite est que l'on ne peut pas l'appliquer lorsque  $\min(n_1, n_2) < k$ , puisqu'alors au moins l'une des deux régressions portant sur les sous-échantillons ne peut plus être calculée. L'article initiateur de Chow (1960) reconnaissait ce problème et proposait un test alternatif pour le traiter. Notre traitement fondé sur la GNR éclaircira la relation entre le test ordinaire (11.03) et le test alternatif. La deuxième limite est que (11.03) n'est compatible qu'avec des modèles de régression linéaire. Il est envisageable de construire l'analogue non linéaire, ce qui nécessite de réaliser deux estimations non linéaires supplémentaires (une pour chaque sous-échantillon). Cependant notre traitement basé sur la GNR offrira un moyen plus simple de manipuler le cas non linéaire.

La troisième limitation relative à (11.03) est que, comme les tests en  $F$  plus conventionnels, il s'agit d'un test qui est valable sous l'hypothèse assez forte que  $E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}$ . Cette hypothèse peut s'avérer assez irréaliste lorsque l'on teste l'égalité de deux ensembles de paramètres, puisque si le vecteur de paramètres  $\beta$  diffère entre les deux régimes, la variance  $\sigma^2$  est aussi sûrement différente. Un certain nombre d'articles a été consacré à cette éventualité, et les nombreux auteurs sont Toyoda (1974), Jayatissa (1977), Schmidt et Sickles (1977), Watt (1979), Honda (1982), Phillips et McCabe (1983), Ali et Silver (1985), Ohtani et Toyoda (1985), Toyoda et Ohtani (1986), Weerahandi (1987), Buse et Dastoor (1989), et Thursby (1992). Tous ces articles considèrent le cas où la variance des aléas est  $\sigma_1^2$  dans le premier régime et  $\sigma_2^2$  dans le second. Une approche qui est souvent plus simple et qui se révèle valable plus souvent consiste à utiliser une statistique de test robuste à l'hétéroscédasticité de forme inconnue (MacKinnon, 1989). C'est plus tard, au cours de la Section 11.6, que nous discuterons d'une régression artificielle qui produit de telles statistiques de test robustes à l'hétéroscédasticité dans tous les cas où la GNR s'applique. Il sera souvent sage de calculer ces tests robustes à l'hétéroscédasticité en plus des tests de Chow ordinaires ou des tests basés sur la GNR, à moins que l'hypothèse d'homoscédasticité soit à l'évidence une hypothèse raisonnable.

Considérons désormais le test de changement de régime dans un modèle de régression non linéaire. Par souci de simplicité, nous supposons que l'échantillon qui doit être partitionné ne doit l'être qu'en deux groupes d'observations; le prolongement de l'analyse au cas d'un nombre de groupes plus important est évident. Nous définissons tout d'abord un vecteur  $\delta \equiv [\delta_1 \cdots \delta_n]^\top$ , en posant  $\delta_t = 0$  si l'observation  $t$  appartient au premier groupe, et  $\delta_t = 1$  si elle appartient au second. Supposons que l'hypothèse nulle soit

$$H_0: y_t = x_t(\beta) + u_t, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I},$$

où, comme d'habitude, les fonctions  $x_t(\beta)$  sont supposées satisfaire les conditions de régularité exposées dans le Chapitre 5. L'hypothèse alternative pourrait s'exprimer comme

$$H_1: y_t = x_t(\beta_1(1 - \delta_t) + \beta_2\delta_t) + u_t, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}.$$

Ainsi, lorsque l'observation  $t$  appartient au groupe 1, de sorte que  $\delta_t = 0$ , la fonction de régression est  $x_t(\beta_1)$ , alors que lorsqu'elle appartient au second groupe, de sorte que  $\delta_t = 1$ , la fonction de régression devient  $x_t(\beta_2)$ .

On peut reformuler l'hypothèse alternative  $H_1$  comme

$$y_t = x_t(\beta_1 + (\beta_2 - \beta_1)\delta_t) + u_t = x_t(\beta_1 + \gamma\delta_t) + u_t,$$

où  $\gamma \equiv \beta_2 - \beta_1$ . Il est clair que  $H_0$  est équivalente à l'hypothèse nulle  $\gamma = \mathbf{0}$ . Puisque cette dernière hypothèse nulle correspond simplement à un ensemble de contraintes de nullité portant sur les paramètres d'une fonction de régression non linéaire, on peut clairement utiliser une régression de Gauss-Newton pour la tester. Cette GNR est

$$y_t - x_t(\hat{\beta}) = \mathbf{X}_t(\hat{\beta})\mathbf{b} + \delta_t \mathbf{X}_t(\hat{\beta})\mathbf{c} + \text{résidu}, \quad (11.04)$$

où  $\hat{\beta}$  désigne les estimations NLS de  $\beta$  sur l'échantillon entier. On peut écrire la GNR (11.04) sous une forme plus compacte comme

$$\hat{\mathbf{u}} = \hat{\mathbf{X}}\mathbf{b} + \delta * \hat{\mathbf{X}}\mathbf{c} + \text{résidus}, \quad (11.05)$$

où  $\hat{\mathbf{u}}$  est composé de l'élément type  $y_t - x_t(\hat{\beta})$ , et  $\hat{\mathbf{X}}$  est composée de l'élément type  $\mathbf{X}_t(\hat{\beta})$ . Le symbole  $*$  désigne ici le **produit direct** de deux matrices. Puisque  $\delta_t \mathbf{X}_{ti}(\hat{\beta})$  est un élément type de  $\delta * \hat{\mathbf{X}}$ ,  $\delta_t * \hat{\mathbf{X}}_t = \hat{\mathbf{X}}_t$  lorsque  $\delta_t = 1$  et  $\delta_t * \hat{\mathbf{X}}_t = \mathbf{0}$  lorsque  $\delta_t = 0$ . Afin d'exécuter le test, il faut simplement estimer le modèle avec l'échantillon entier, et régresser les résidus de cette estimation sur la matrice des dérivées  $\hat{\mathbf{X}}$  et sur la matrice dont les lignes qui correspondent aux observations du groupe 1 sont composées de zéros. Il est inutile d'ordonner les données. Comme d'habitude, on dispose de plusieurs statistiques de test asymptotiquement valables, la meilleure étant sûrement la statistique  $F$  ordinaire pour l'hypothèse nulle  $\mathbf{c} = \mathbf{0}$ . Dans le cas le plus courant où  $k$  est plus petit que  $\min(n_1, n_2)$ , cette statistique de test aura  $k$  degrés de liberté au numérateur et  $n - 2k$  degrés de liberté au dénominateur.

Notons que la SSR de la régression (11.05) est égale à la SSR de la GNR

$$\hat{\mathbf{u}} = \hat{\mathbf{X}}\mathbf{b} + \text{résidus} \quad (11.06)$$

exécutée sur les observations 1 à  $n_1$  plus la SSR de la régression (11.05) exécutée sur les observations  $n_1 + 1$  à  $n$ . Cette SSR correspond à la SSR non contrainte pour le test en  $F$  de  $\mathbf{c} = \mathbf{0}$  dans (11.05). La SSR contrainte pour ce test est tout simplement la SSR de (11.06) calculée sur les  $n$  observations, qui est identique à la SSR de l'estimation non linéaire de l'hypothèse nulle  $H_0$ . Ainsi le test de Chow ordinaire pour la GNR (11.06) sera numériquement identique au test en  $F$  de  $\mathbf{c} = \mathbf{0}$  dans (11.05). Cette propriété fournit le moyen le plus aisé de calculer la statistique de test.

Comme nous l'avons mentionné plus haut, le test de Chow ordinaire (11.03) ne s'applique pas lorsque  $\min(n_1, n_2) < k$ . L'usage de la structure de la GNR montre clairement pourquoi c'est le cas. Sans perte de généralité, puisque la numérotation des deux groupes d'observations est arbitraire, supposons que  $n_2 < k$  et  $n_1 > k$ . Alors, la matrice  $\delta * \hat{\mathbf{X}}$ , qui a  $k$  colonnes, possédera  $n_2 < k$  lignes qui ne sont pas uniquement composées de zéros et par conséquent aura un rang au plus égal à  $n_2$ . Ainsi, lorsque l'on estime l'équation (11.05), au plus  $n_2$  éléments de  $\mathbf{c}$  seront identifiés, et les résidus correspondant à toutes les observations du second groupe seront nuls. Par conséquent le nombre de degrés de liberté au numérateur de la statistique  $F$  est au plus égal à  $n_2$ . En réalité, il sera égal au rang de  $[\hat{\mathbf{X}} \quad \delta * \hat{\mathbf{X}}]$  moins celui de  $\hat{\mathbf{X}}$ , ce qui pourrait donner un résultat inférieur à  $n_2$  dans certains cas. Le nombre de degrés de liberté pour le dénominateur correspondra au nombre d'observations pour lesquelles les résidus de (11.05) sont nuls, c'est-à-dire  $n_1$ , moins le nombre de régresseurs associés à ces mêmes observations, c'est-à-dire  $k$ , soit un total de  $n_1 - k$ . Ainsi nous pouvons utiliser la GNR que  $\min(n_1, n_2) < k$  soit vérifié ou pas, à condition d'utiliser le nombre de degrés de liberté adéquat pour le numérateur et le dénominateur du test en  $F$ .

Il devrait être clair que lorsque  $x_t(\boldsymbol{\beta}) = \mathbf{X}_t\boldsymbol{\beta}$  et  $\min(n_1, n_2) > k$ , le test en  $F$  basé sur la GNR (11.05) est *numériquement identique* au test de Chow (11.03). Cette propriété découle du fait que la somme des résidus au carré de (11.05) sera alors égale à  $\text{SSR}_1 + \text{SSR}_2$ , la somme des deux SSR provenant des estimations séparées de la régression sur les deux groupes d'observations. La démonstration de l'identité numérique entre le test "alternatif" de Chow (1960) et le test correspondant basé sur la GNR (qui a  $n_2$  et  $n_1 - k$  degrés de liberté dans les cas réguliers) lorsque  $x_t(\boldsymbol{\beta}) = \mathbf{X}_t\boldsymbol{\beta}$  et  $\min(n_1, n_2) < k$ , serait un bon exercice.

Quelquefois, nous voudrions tester l'égalité d'un sous-ensemble de paramètres du modèle sur deux sous-échantillons, plutôt que l'ensemble entier. Il est très aisé de modifier les tests dont nous venons de discuter pour appréhender cette situation. Les hypothèses nulle et alternative peuvent désormais s'écrire

$$H_0: y_t = x_t(\boldsymbol{\alpha}, \boldsymbol{\beta}) + u_t, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}, \quad \text{et} \quad (11.07)$$

$$H_1: y_t = x_t(\boldsymbol{\alpha}, \boldsymbol{\beta}_1(1 - \delta_t) + \boldsymbol{\beta}_2\delta_t) + u_t, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I},$$

où  $\boldsymbol{\alpha}$  est un vecteur de  $l$  paramètres qui sont supposés être identiques dans les deux sous-échantillons, et  $\boldsymbol{\beta}$  est un vecteur à  $m$  composantes dont on veut tester la constance. La GNR est alors

$$\hat{\mathbf{u}} = \hat{\mathbf{X}}_{\boldsymbol{\alpha}}\mathbf{a} + \hat{\mathbf{X}}_{\boldsymbol{\beta}}\mathbf{b} + \delta * \hat{\mathbf{X}}_{\boldsymbol{\beta}}\mathbf{c} + \text{résidus},$$

où  $\hat{\mathbf{X}}_{\boldsymbol{\alpha}}$  est une matrice de dimension  $n \times l$ , dont l'élément type est la dérivée partielle  $\partial x_t(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \alpha_i$ , évaluée en  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ , les estimations de (11.07); et où

$\hat{\mathbf{X}}_{\beta}$  est une matrice de dimension  $n \times m$  dont l'élément type est  $\partial x_t(\alpha, \beta) / \partial \beta_j$ , également évalué en  $(\hat{\alpha}, \hat{\beta})$ . A condition que  $m$  soit inférieur à  $\min(n_1, n_2)$ , la statistique de test aura  $m$  et  $n - l - 2m$  degrés de liberté. Même dans le cas où  $x_t(\alpha, \beta)$  est linéaire, il n'est pas encore possible de calculer un test à la manière du test de Chow classique (11.03). Parce que le vecteur de paramètres  $\alpha$  est supposé être constant sur les deux sous-échantillons, on ne peut pas obtenir la SSR non contrainte grâce à une estimation séparée des deux sous-échantillons.

La discussion précédente se focalisait sur le contexte de l'estimation par moindres carrés. Lorsque l'on effectue une estimation par variables instrumentales, il y a une légère complication relative au choix des instruments à utiliser lors de l'estimation du modèle nul et du modèle alternatif. D'après les résultats de la Section 7.7, l'équivalent IV de (11.05) est

$$\tilde{\mathbf{u}} = \mathbf{P}_W \tilde{\mathbf{X}} \mathbf{b} + \mathbf{P}_W \delta^* \tilde{\mathbf{X}} \mathbf{c} + \text{résidus}, \quad (11.08)$$

où  $\tilde{\mathbf{u}}$  et  $\tilde{\mathbf{X}}$  sont évalués avec les estimations  $\tilde{\beta}$  des IV (généralisés) sous l'hypothèse nulle. Comme d'habitude, de nombreuses statistiques de test sont disponibles.

Bien que l'estimation de la régression (11.08) semble immédiate, un problème demeure. En faisant usage simplement de la matrice des instruments  $\mathbf{W}$  qui a permis l'estimation du modèle originel, il est fort possible que la matrice  $[\mathbf{P}_W \tilde{\mathbf{X}} \quad \mathbf{P}_W \delta^* \tilde{\mathbf{X}}]$  ne soit pas de plein rang. Pour estimer le modèle contraint,  $\mathbf{W}$  doit avoir au moins  $k$  colonnes, alors que pour effectuer la régression (11.08) elle doit en avoir au moins  $2k$ . Si  $\mathbf{W}$  possède moins de  $2k$  colonnes, la statistique de test aura un nombre de degrés de liberté inférieur à  $k$ , et testera en réalité  $H_0$  contre une hypothèse alternative moins générale que  $H_1$ . Une solution immédiate consiste à doubler le nombre des instruments par l'usage de la matrice

$$\mathbf{W}^* \equiv \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix} \quad (11.09)$$

au lieu de  $\mathbf{W}$  dans la GNR (11.08). Cela permet aux relations entre les régresseurs endogènes et les instruments de différer dans les deux sous-échantillons, ce qui paraît raisonnable. Même s'il faut utiliser un test LM, c'est-à-dire un test basé sur la somme des carrés expliqués de la régression (11.08), il faut être prudent et utiliser  $\mathbf{W}^*$  pour l'estimation du modèle *contraint*. Toutefois, ainsi que nous en avons discuté au cours de la Section 7.7, cela n'est pas nécessaire lorsque l'on utilise un test  $C(\alpha)$ , c'est-à-dire un test pseudo- $F$  pour  $\mathbf{c} = \mathbf{0}$  dans la régression (11.08).

Il est sans doute utile d'énumérer les étapes qu'il faut franchir si l'on désire tester  $H_0$  contre  $H_1$  avec des estimations IV:

- (i) Estimer le modèle  $H_0$  grâce à une matrice  $\mathbf{W}$  adéquate composée d'au moins  $k$  instruments, et de préférence de plus que  $k$ , comprenant toutes les variables exogènes et prédéterminées dans les fonctions de régression.

- (ii) Elaborer une nouvelle matrice d'instruments  $\mathbf{W}^*$  comme dans (11.09). Puis, afin d'obtenir la SSR contrainte, estimer la GNR

$$\tilde{\mathbf{u}} = \mathbf{P}_{\mathbf{W}^*} \tilde{\mathbf{X}} \mathbf{b} + \text{résidus}$$

sur l'échantillon entier, avec  $\tilde{\mathbf{u}}$  et  $\tilde{\mathbf{X}}$  évalués avec les estimations IV obtenues en l'étape (i).

- (iii) Pour obtenir la SSR non contrainte, exécuter la GNR

$$\tilde{\mathbf{u}}_j = \mathbf{P}_{\mathbf{W}_j} \tilde{\mathbf{X}}_j \mathbf{b} + \text{résidus}$$

sur les deux sous-échantillons séparément et additionner les deux sommes des résidus au carré. Ici  $\tilde{\mathbf{u}}_j$ ,  $\mathbf{W}_j$ ,  $\tilde{\mathbf{X}}_j$  désignent le sous-vecteur  $\tilde{\mathbf{u}}$  et les sous-matrices  $\mathbf{W}$ , et  $\tilde{\mathbf{X}}$  qui correspondent aux deux sous-échantillons.

- (iv) Calculer un test  $C(\alpha)$ , ou un test pseudo- $F$  basé sur les résidus des régressions obtenues en (ii) et (iii), suivant la procédure décrite dans la Section 7.7.

Une procédure alternative consisterait à estimer à la fois le modèle contraint et le modèle non contraint avec  $\mathbf{W}^*$  comme matrice d'instruments, mais cela serait considérablement plus difficile dans le cas non linéaire. Pour le modèle non contraint, cela impliquerait l'estimation par IV de chaque sous-échantillon séparément, en utilisant  $\mathbf{W}_j$  pour chaque sous-échantillon  $j$ . Alors il serait possible de calculer n'importe quelle statistique de tests basée sur les estimations contraintes et non contraintes dont nous avons discuté à la Section 7.7.

La littérature consacrée aux tests de changement de régime est importante, et trop étendue pour que nous en discutions dans cette section. Un certain nombre de contributions récentes à ce domaine, ainsi qu'une bibliographie utile se retrouvent chez Krämer (1989).

### 11.3 TESTS DE MODÈLES DE RÉGRESSION NON EMBOÎTÉS

Tous les tests que nous avons étudiés jusqu'à présent impliquaient des **modèles emboîtés**. Cela signifie que le modèle que l'on teste, représenté par l'hypothèse nulle, est un cas particulier du modèle alternatif contre lequel il est testé. Par exemple, un modèle de régression avec des aléas non autocorrélés est un cas particulier du modèle alternatif dont les aléas sont AR(1), et le modèle dont les coefficients sont constants sur un échantillon entier est un cas particulier du modèle alternatif dont les coefficients varient entre deux sous-échantillons. Bien que des modèles alternatifs emboîtés apparaissent fréquemment, il existe aussi de nombreux cas où les deux modèles (ou davantage) qui s'excluent ne sont pas emboîtés. La littérature consacrée aux **tests d'hypothèses non emboîtées** a rendu envisageable la manipulation de telles situations au sein de la structure de la régression de Gauss-Newton.

Bien que notre traitement se fasse dans un contexte de régressions artificielles, ce n'est pas le cas de la majeure partie des premiers articles consacrés aux tests des hypothèses non emboîtées. Les références classiques sont les deux articles de Cox (1961, 1962) et les deux articles d'Atkinson (1969, 1970). Les idées de base de Cox furent adaptées aux modèles de régression linéaire par Pesaran (1974) et aux modèles de régression non linéaire par Pesaran et Deaton (1978). L'approche par la régression artificielle est due à Davidson et MacKinnon (1981a).

Supposons que deux théories économiques concurrentes (ou deux concrétisations d'un même modèle théorique de base) qui prétendent toutes deux avoir un pouvoir explicatif sur la même variable dépendante, produisent les deux modèles de régression non linéaire:

$$\begin{aligned} H_1: \quad \mathbf{y} &= \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}_1, & E(\mathbf{u}_1 \mathbf{u}_1^\top) &= \sigma_1^2 \mathbf{I}, \text{ et} \\ H_2: \quad \mathbf{y} &= \mathbf{z}(\boldsymbol{\gamma}) + \mathbf{u}_2, & E(\mathbf{u}_2 \mathbf{u}_2^\top) &= \sigma_2^2 \mathbf{I}, \end{aligned}$$

où  $\boldsymbol{\beta}$  et  $\boldsymbol{\gamma}$  sont des vecteurs de dimensions respectives égales à  $k_1$  et  $k_2$ . Ces modèles sont dits **non emboîtés** s'il est en général impossible de trouver des contraintes sur  $\boldsymbol{\beta}$  telles que, pour un  $\boldsymbol{\gamma}$  donné,  $\mathbf{x}(\boldsymbol{\beta})$  égale  $\mathbf{z}(\boldsymbol{\gamma})$ , et impossible de trouver des contraintes sur  $\boldsymbol{\gamma}$  telles que, pour un  $\boldsymbol{\beta}$  donné,  $\mathbf{z}(\boldsymbol{\gamma})$  égale  $\mathbf{x}(\boldsymbol{\beta})$ . Ainsi, il ne doit exister aucune application, disons  $\mathbf{g}$ , définie sur l'espace paramétrique entier sur lequel  $\boldsymbol{\gamma}$  est défini, telle que  $\mathbf{z}(\boldsymbol{\gamma}) = \mathbf{x}(\mathbf{g}(\boldsymbol{\gamma}))$ . De façon similaire, il ne doit exister aucune application  $\mathbf{h}$  telle que  $\mathbf{x}(\boldsymbol{\beta}) = \mathbf{z}(\mathbf{h}(\boldsymbol{\beta}))$ .

Il est nécessaire, dans le cas des modèles de régression linéaire, que chacune des deux fonctions de régression comprenne au moins un régresseur que l'on ne trouve pas dans l'autre. Par exemple, les deux fonctions de régression suivantes sont non emboîtées:

$$x_t(\boldsymbol{\beta}) = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} \quad \text{et} \quad (11.10)$$

$$z_t(\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 X_{t1} + \gamma_3 X_{t3}. \quad (11.11)$$

Toutefois, si l'on ajoutait  $X_{t2}$  à (11.11) pour aboutir à la nouvelle fonction de régression

$$z_t^*(\boldsymbol{\gamma}) = \gamma_0 + \gamma_1 X_{t1} + \gamma_2 X_{t2} + \gamma_3 X_{t3}, \quad (11.12)$$

(11.10) serait emboîtée dans (11.12), puisqu'en annulant  $\gamma_3$  on rendrait (11.12) équivalente à (11.10).

Les tests non emboîtés les plus faciles à mettre en œuvre sont ceux basés sur l'**emboîtement artificiel**. L'idée générale est qu'il faut combiner les deux fonctions de régression concurrentes pour en construire une plus générale et tester un ou les deux modèles initiaux contre cette nouvelle fonction. Considérons le modèle artificiel composé

$$H_C: \quad \mathbf{y} = (1 - \alpha)\mathbf{x}(\boldsymbol{\beta}) + \alpha\mathbf{z}(\boldsymbol{\gamma}) + \mathbf{u}, \quad (11.13)$$

où  $\alpha$  est un paramètre qui a été introduit afin d'emboîter  $H_1$  et  $H_2$  au sein de  $H_C$ ; lorsque  $\alpha = 0$ ,  $H_C$  se réduit à  $H_1$ , et lorsque  $\alpha = 1$ ,  $H_C$  se réduit à  $H_2$ . Le problème réside dans le fait que, dans la plupart des cas, il ne sera pas possible d'estimer le modèle artificiel (11.13), parce que les paramètres  $\alpha$ ,  $\beta$ , et  $\gamma$  ne seront pas identifiables séparément. Par exemple, en conservant (11.10) et (11.11),  $H_C$  aura sept paramètres en tout (les trois  $\beta_i$ , les trois  $\gamma_i$ , et  $\alpha$ ) mais ne pourra en identifier et en estimer en réalité que quatre (la constante et les trois coefficients de  $X_1$ ,  $X_2$ , et  $X_3$ ).

Une solution à ce problème dont l'initiative revient à Davidson et MacKinnon (1981a), consiste à remplacer  $H_C$  par un modèle où les paramètres inconnus du modèle qui *n'est pas* testé sont remplacés par des estimations de ces paramètres qui seraient convergentes si le DGP appartenait en réalité au modèle dans lequel ils sont définis. Supposons que l'on veuille tester  $H_1$ . Alors il faut remplacer  $\gamma$  dans (11.13) par une estimation convergente sous  $H_2$ . Il y a plusieurs manières d'y parvenir, puisqu'il existe plusieurs manières d'obtenir des estimations convergentes de  $\gamma$ , mais la plus simple et celle qui a les propriétés asymptotiques les plus intéressantes consiste à prendre  $\hat{\gamma}$ , l'estimation NLS de  $\gamma$ . Ainsi  $H_C$  devient

$$H'_C : \mathbf{y} = (1 - \alpha)\mathbf{x}(\beta) + \alpha\hat{\mathbf{z}} + \mathbf{u}, \quad (11.14)$$

où  $\hat{\mathbf{z}} \equiv \mathbf{z}(\hat{\gamma})$ . Le nouveau modèle composé  $H'_C$  ne possède plus que  $k_1 + 1$  paramètres à estimer, soit un de plus que  $H_1$ . A condition que  $H_1$  et  $H_2$  soient véritablement non emboîtées et que  $H_1$  soit identifiée asymptotiquement, à la fois  $\alpha$  et  $\beta$  doivent être identifiables asymptotiquement. Il est alors possible de tester  $H_1$  en testant l'hypothèse nulle  $\alpha = 0$ , grâce à n'importe quel test habituel. Davidson et MacKinnon (1981a) ont suggéré deux éventualités pour ce genre de test. Le **test en J** utilise le  $t$  de Student pour  $\alpha = 0$  à partir de l'estimation non linéaire de (11.14). Il porte le nom de test en  $J$  parce que  $\alpha$  et  $\beta$  sont estimés *conjointement*. Etant donné que cela pourrait être difficile lorsque  $\mathbf{x}(\beta)$  est non linéaire, les deux auteurs ont proposé une procédure alternative, appelée **test en P**. Il faut utiliser le  $t$  de Student pour  $\alpha = 0$  dans la régression de Gauss-Newton

$$\mathbf{y} - \hat{\mathbf{x}} = \hat{\mathbf{X}}\mathbf{b} + a(\hat{\mathbf{z}} - \hat{\mathbf{x}}) + \text{résidus}, \quad (11.15)$$

où  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\beta})$  et  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\beta})$ , où  $\mathbf{X}(\beta)$  désigne la matrice de dimension  $n \times k_1$  des dérivées de  $\mathbf{x}(\beta)$  par rapport à  $\beta$  et  $\hat{\beta}$  désigne l'estimation NLS de  $\beta$  sous  $H_1$ . Le régresseur de test  $\hat{\mathbf{z}} - \hat{\mathbf{x}}$  est obtenu, comme d'habitude, en prenant la dérivée partielle de la fonction de régression du modèle  $H'_C$  par rapport à  $\alpha$  et en l'évaluant en  $\alpha = 0$ ,  $\beta = \hat{\beta}$ .<sup>3</sup>

<sup>3</sup> Remarquons que le test en  $P$  pourrait également être utilisé dans des situations où l'on dispose des estimations  $\beta$  et  $\gamma$  convergentes au taux  $n^{1/2}$  mais pas des estimations par moindres carrés. C'est une simple applications des résultats de la Section 6.7



Au vu des résultats généraux sur les régressions de Gauss-Newton du Chapitre 6, il est évident que les tests en  $J$  et en  $P$  sont asymptotiquement équivalents sous  $H_1$ . Ainsi, si l'un de ces tests est valable asymptotiquement, l'autre doit l'être aussi. Cependant, il n'est pas immédiatement évident que l'un des tests soit en réalité valable, puisque  $\hat{\mathbf{z}}$ , qui dépend de  $\mathbf{y}$ , apparaît dans le membre de droite de (11.14). L'intuition de la validité asymptotique est toutefois assez simple. A condition que, sous  $H_1$ , le vecteur  $\hat{\gamma}$  converge asymptotiquement vers un quelconque vecteur constant, disons  $\gamma_1$ , alors le vecteur  $\hat{\mathbf{z}} \equiv \mathbf{z}(\hat{\gamma})$  doit également converger vers un vecteur  $\mathbf{z}(\gamma_1)$ . Il est donc asymptotiquement valable de traiter le vecteur  $\hat{\mathbf{z}}$  comme s'il s'agissait d'un vecteur d'observations sur une variable prédéterminée.

Lorsque  $\mathbf{x}(\beta) = \mathbf{X}\beta$ , le modèle soumis au test est linéaire. Dans ce cas, la régression du test en  $J$  (11.14) doit produire exactement les mêmes résultats que la régression du test en  $P$  (11.15). Parce que  $\hat{\mathbf{x}} = \mathbf{X}\hat{\beta}$ , il est clair que  $\mathcal{S}(\mathbf{X}, \hat{\mathbf{z}})$  est exactement identique à  $\mathcal{S}(\mathbf{X}, \hat{\mathbf{z}} - \hat{\mathbf{x}})$ . Ainsi, les deux régressions doivent posséder le même pouvoir explicatif et par conséquent doivent produire des statistiques de test identiques.

Il est aussi juste de tester  $H_2$  contre  $H_C$  que de tester  $H_1$  contre  $H_C$ , et la régression artificielle est essentiellement la même que la précédente, mais  $H_1$  prend maintenant la place de  $H_2$  et vice versa. Ainsi la régression équivalente à (11.14) pour le test en  $J$  est

$$\mathbf{y} = (1 - \phi)\mathbf{z}(\gamma) + \phi\hat{\mathbf{x}} + \mathbf{u},$$

et la régression équivalente à (11.15) pour le test en  $P$  est

$$\mathbf{y} - \hat{\mathbf{z}} = \hat{\mathbf{Z}}\mathbf{c} + p(\hat{\mathbf{x}} - \hat{\mathbf{z}}) + \text{résidus}.$$

Remarquons qu'il *ne serait pas* pertinent d'utiliser (11.14) ou même (11.15) pour tester  $H_2$ .

Lorsque l'on effectue un couple de tests de modèles non emboîtés, il y a quatre résultats possibles, puisque aussi bien  $H_1$  que  $H_2$  peut être rejetée ou non. Si, par exemple,  $H_1$  est rejetée et  $H_2$  ne l'est pas, alors il paraît raisonnable de conserver  $H_2$  comme le modèle le plus adéquat. Mais il est aussi possible que les deux modèles soient rejetés ou qu'aucun d'entre eux ne le soit. Lorsque l'on rejette les deux modèles, il nous faut conclure qu'aucun n'est satisfaisant, ce qui est une éventualité peu plaisante mais qui nous stimulera pour développer des modèles plus complets. Lorsqu'aucun n'est rejeté, il nous faut conclure que les modèles s'ajustent aux données apparemment avec la même qualité et qu'aucun n'assure avec évidence que l'autre est mal spécifié. Sans doute, soit les deux modèles sont très similaires, soit l'ensemble des données porte peu d'information. Le fait qu'une paire de tests d'hypothèses non emboîtées ne nous permette pas en général de choisir un modèle plutôt qu'un autre peut être considéré comme une déficience de ces tests. C'est le cas si l'on interprète mal leur nature. Les tests d'hypothèses

non emboîtées sont des tests de spécification, et puisqu'il n'y a presque jamais aucune raison a priori de croire qu'un des modèles a en réalité généré les données, il est pertinent que les tests non emboîtés, tout comme les autres tests de spécification d'un modèle, nous enseignent qu'aucun des modèles n'est compatible aux données.

Il est important d'insister sur le fait que l'objet des tests non emboîtés *n'est pas* de choisir le "meilleur" modèle parmi un ensemble fixé de modèles. Ceci constitue l'objet d'un pan entièrement différent de la littérature économétrique, qui traite les critères d'une **sélection de modèle**. Nous n'entamerons pas de débat sur la littérature assez importante consacrée à la sélection de modèle à travers cet ouvrage. Deux études utiles sont dues à Amemiya (1980) et Leamer (1983), et un intéressant article récent a été écrit par Pollak et Wales (1991).

Il est intéressant d'examiner plus en détail le cas où les deux modèles sont linéaires, c'est-à-dire  $\mathbf{x}(\beta) = \mathbf{X}\beta$  et  $\mathbf{z}(\gamma) = \mathbf{Z}\gamma$ . Cela nous donnera l'occasion de comprendre pourquoi les tests en  $J$  et en  $P$  (qui sont identiques dans ce cas précis) sont asymptotiquement valables et aussi pourquoi ces tests peuvent ne pas toujours être performants lorsque les échantillons sont finis. La régression du test en  $J$  pour tester  $H_1$  contre  $H_2$  est

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \alpha\mathbf{P}_Z\mathbf{y} + \text{résidus}, \quad (11.16)$$

où  $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$  et  $\mathbf{b} = (1 - \alpha)\beta$ . L'usage du Théorème FWL nous permet de voir que l'estimation de  $\alpha$  dans (11.16) sera identique à l'estimation de  $\alpha$  dans la régression

$$\mathbf{M}_X\mathbf{y} = \alpha\mathbf{M}_X\mathbf{P}_Z\mathbf{y} + \text{résidus}. \quad (11.17)$$

Ainsi, si  $\hat{\sigma}$  désigne l'estimation OLS de  $\sigma$  à partir de (11.17), le  $t$  de Student de  $\alpha = 0$  sera

$$\frac{\mathbf{y}^\top\mathbf{P}_Z\mathbf{M}_X\mathbf{y}}{\hat{\sigma}(\mathbf{y}^\top\mathbf{P}_Z\mathbf{M}_X\mathbf{P}_Z\mathbf{y})^{1/2}}. \quad (11.18)$$

Tout d'abord, remarquons que lorsqu'une seule colonne de  $\mathbf{Z}$ , disons  $\mathbf{Z}_1$ , n'appartient pas à  $\mathcal{S}(\mathbf{X})$ , alors

$$\mathcal{S}(\mathbf{X}, \mathbf{P}_Z\mathbf{y}) = \mathcal{S}(\mathbf{X}, \mathbf{Z}) = \mathcal{S}(\mathbf{X}, \mathbf{Z}_1).$$

Par conséquent, la régression du test en  $J$  (11.16) doit produire exactement la même SSR que la régression

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \delta\mathbf{Z}_1 + \text{résidus}. \quad (11.19)$$

Ainsi, dans ce cas particulier, le test en  $J$  est égal en valeur absolue au  $t$  de Student de l'estimation de  $\delta$  dans (11.19).

Lorsque deux ou plus de deux colonnes de  $\mathbf{Z}$  n'appartiennent pas à  $\mathcal{S}(\mathbf{X})$ , ce résultat particulier n'est plus valable. Si les données sont réellement générées par  $H_1$ , nous pouvons remplacer  $\mathbf{y}$  dans le numérateur de (11.18) par  $\mathbf{X}\beta + \mathbf{u}$ . Puisque  $\mathbf{M}_X \mathbf{X}\beta = \mathbf{0}$ , ce numérateur devient

$$\beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{u} + \mathbf{u}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{u}. \quad (11.20)$$

Les deux termes de (11.20) sont de natures différentes. Le premier terme est une somme pondérée des éléments du vecteur  $\mathbf{u}$ , dont chacun est d'espérance nulle. Ainsi, sous les conditions de régularité adéquates, il est aisé de voir que

$$n^{-1/2} \beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{u} \stackrel{a}{\sim} N\left(\mathbf{0}, \operatorname{plim}_{n \rightarrow \infty} (n^{-1} \sigma_1^2 \beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X} \beta)\right).$$

Le premier terme est donc  $O(n^{1/2})$ . Par contraste, le second terme est  $O(1)$ , puisque

$$\begin{aligned} \operatorname{plim}_{n \rightarrow \infty} (\mathbf{u}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{u}) &= \operatorname{plim}_{n \rightarrow \infty} (\mathbf{u}^\top \mathbf{P}_Z \mathbf{u} - \mathbf{u}^\top \mathbf{P}_Z \mathbf{P}_X \mathbf{u}) \\ &= \sigma_1^2 k_2 - \sigma_1^2 \operatorname{Tr} \left( \lim_{n \rightarrow \infty} \mathbf{P}_Z \mathbf{P}_X \right), \end{aligned}$$

et la trace de  $\mathbf{P}_Z \mathbf{P}_X$  est  $O(1)$ . Ainsi, seul le premier terme de (11.20) garde un intérêt asymptotiquement.

De façon similaire, sous  $H_1$ , le facteur entre parenthèses dans le dénominateur de (11.18) est égal à

$$\beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X} \beta + 2\beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{u} + \mathbf{u}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{u}. \quad (11.21)$$

Avec des arguments comparables à ceux utilisés pour le numérateur, on pourrait montrer que le premier des trois termes de (11.21) est  $O(n)$ , que le deuxième est  $O(n^{1/2})$ , et que le troisième est  $O(1)$ . Ainsi, sous  $H_1$ , la statistique de test (11.18) tend asymptotiquement vers la variable aléatoire

$$\frac{\beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{u}}{\sigma_1 (\beta^\top \mathbf{X}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{P}_Z \mathbf{X} \beta)^{1/2}},$$

dont on peut montrer qu'elle suit asymptotiquement une  $N(0, 1)$ .

Cette analyse ne montre pas seulement pourquoi les tests en  $J$  et en  $P$  sont valables asymptotiquement mais elle indique également pourquoi ils peuvent mal se comporter avec des échantillons finis. Lorsque la taille de l'échantillon est faible ou que  $\mathbf{Z}$  contient plusieurs régresseurs qui n'appartiennent pas à  $\mathcal{S}(\mathbf{X})$ , la quantité  $\mathbf{u}^\top \mathbf{P}_Z \mathbf{M}_X \mathbf{u}$ , qui est négligeable asymptotiquement, peut en réalité être importante et positive. Par conséquent, dans de telles circonstances, la statistique de test du test en  $J$  (11.18) peut être d'espérance substantiellement supérieure à zéro.

De nombreux moyens de réduire ou d'éliminer ce biais ont été proposés. Le plus simple d'entre eux, dont l'initiative revient à Fisher et McAleer (1981) et qui a été étudié plus tard par Godfrey (1983), consiste à remplacer  $\hat{\gamma}$  dans les régressions du test en  $J$  et du test en  $P$  par  $\tilde{\gamma}$ , qui est l'estimation de  $\gamma$  obtenue par la minimisation de

$$(\hat{\mathbf{x}} - \mathbf{z}(\gamma))^{\top}(\hat{\mathbf{x}} - \mathbf{z}(\gamma)).$$

Ainsi  $\tilde{\gamma}$  est l'estimation NLS de  $\gamma$  obtenue lorsque l'on utilise les valeurs ajustées  $\hat{\mathbf{x}}$  au lieu de la variable dépendante  $\mathbf{y}$ . Dans le cas linéaire, cela signifie que la régression du test en  $J$  (11.16) est remplacée par la régression

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \alpha\mathbf{P}_Z\mathbf{P}_X\mathbf{y} + \text{résidus}. \quad (11.22)$$

Cette régression produit ce que l'on appelle le **test en  $J_A$**  parce que Fisher et McAleer en ont attribué l'initiative à Atkinson (1970). En utilisant un résultat de Milliken et Graybill (1970), Godfrey (1983) montra que le  $t$  de Student de l'estimation de  $\alpha$  à partir de (11.22) obéit réellement à la distribution de Student avec des échantillons finis sous les conditions habituelles qui font que le  $t$  de Student suit cette distribution ( $\mathbf{u}$  normalement distribué, indépendance de  $\mathbf{y}$  à l'égard de  $\mathbf{X}$  et de  $\mathbf{Z}$ ). L'intuition de ce résultat est relativement simple. Le vecteur de valeurs ajustées  $\mathbf{P}_X\mathbf{y}$  ne contient que la partie de  $\mathbf{y}$  qui appartient à  $\mathcal{S}(\mathbf{X})$ . Il doit donc être indépendant de  $\mathbf{M}_X\mathbf{y}$ , qui est ce que seraient les résidus de (11.22) si  $\alpha = 0$ . Par conséquent, il est possible de traiter  $\mathbf{P}_Z\mathbf{P}_X\mathbf{y}$  (ou n'importe quel autre régresseur qui ne dépend de  $\mathbf{y}$  qu'à travers  $\mathbf{P}_X\mathbf{y}$ ) comme s'il s'agissait d'un régresseur fixe.<sup>4</sup> Le **test en  $P_A$**  est au test en  $P$  ce que le test en  $J_A$  est au test en  $J$ .

Malheureusement, les tests en  $J_A$  et en  $P_A$  sont dans de nombreuses circonstances beaucoup moins puissants que les tests en  $J$  et en  $P$  ordinaires; consulter Davidson et MacKinnon (1982) et Godfrey et Pesaran (1983). Alors si, par exemple, le test  $J$  rejette l'hypothèse nulle alors que le test en  $J_A$  ne le fait pas, il est difficile de savoir si c'est parce que le premier est très enclin à commettre une erreur de première espèce ou parce que le second est au contraire très enclin à commettre une erreur de deuxième espèce.

Une seconde approche consiste à estimer l'espérance de  $\mathbf{u}^{\top}\mathbf{M}_X\mathbf{P}_Z\mathbf{u}$ , à la soustraire de  $\mathbf{y}^{\top}\mathbf{M}_X\mathbf{P}_Z\mathbf{y}$ , et à diviser la quantité qui en résulte par une estimation de la racine carrée de la variance afin d'obtenir une statistique de test qui obéirait asymptotiquement à une  $N(0, 1)$ . Cette approche a été proposée par Godfrey et Pesaran (1983) sous une forme quelque peu compliquée; on trouvera une version plus simple dans le "Reply" de MacKinnon (1983). Cette seconde approche est beaucoup plus difficile à mettre en œuvre que le test en  $J_A$ , puisqu'elle implique des calculs matriciels qui ne peuvent pas être

<sup>4</sup> Avec le même argument, le test RESET dont nous avons discuté à la Section 3.5 est exact avec des échantillons finis toutes les fois qu'un test en  $t$  ordinaire l'est.

réalisés avec une succession de régressions, et elle ne produit pas un test exact. De plus elle nécessite l'hypothèse de normalité. Cependant, il semble qu'elle génère un test dont les propriétés en échantillons finis sont bien meilleures que celles du test en  $J$  sous l'hypothèse nulle, et dont la puissance, du moins dans certains cas, sera supérieure à celle du test en  $J_A$ .

Le vecteur  $\tilde{\gamma}$  est intéressant de plein droit. Le test de Cox originel utilisait le fait que, sous  $H_1$ ,

$$\text{plim}_{n \rightarrow \infty}(\tilde{\gamma}) = \text{plim}_{n \rightarrow \infty}(\hat{\gamma}).$$

Il est possible d'élaborer un test basé directement sur la différence entre  $\hat{\gamma}$  et  $\tilde{\gamma}$ . Un tel test, proposé pour la première fois par Dastoor (1983) et développé plus tard par Mizon et Richard (1986), détermine si oui ou non la valeur de  $\gamma$  prédite par le modèle  $H_1$  (c'est-à-dire  $\tilde{\gamma}$ ) est la même que la valeur obtenue par l'estimation directe de  $H_2$  (c'est-à-dire  $\hat{\gamma}$ ). On appelle ces tests les **tests d'enveloppement** parce que si  $H_1$  explique effectivement la réalisation de  $H_2$ , on peut dire qu'elle l'enveloppe; voir Mizon (1984). Le principe sur lequel ils reposent est quelquefois appelé **principe d'enveloppement**.

Il y a quelques difficultés pratiques avec les modèles de régression non linéaire, et par conséquent nous ne discuterons pas de ces tests dans cet ouvrage. Toutefois, dans le cas linéaire, le test est à la fois simple et attrayant. Lorsque les deux modèles sont linéaires, les deux estimations de  $\gamma$  sont

$$\begin{aligned}\hat{\gamma} &= (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad \text{et} \\ \tilde{\gamma} &= (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{P}_X \mathbf{y}.\end{aligned}$$

Ainsi la différence entre les deux est

$$(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} - (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{P}_X \mathbf{y} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{M}_X \mathbf{y}. \quad (11.23)$$

Le facteur  $(\mathbf{Z}^\top \mathbf{Z})^{-1}$  est à l'évidence sans pertinence dans la construction de n'importe quelle statistique de test. Le vecteur  $\mathbf{Z}^\top \mathbf{M}_X \mathbf{y}$  sera en général composé d'éléments nuls, chacun de ces éléments correspondant à chaque colonne de  $\mathbf{Z}$  qui appartient à  $\mathcal{S}(\mathbf{X})$ . Posons  $\mathbf{Z}^*$  la matrice composée des colonnes restantes de  $\mathbf{Z}$ . Alors il devrait être clair à partir de (11.23) que ce que nous voulons réellement tester, c'est si le vecteur  $\mathbf{Z}^{*\top} \mathbf{M}_X \mathbf{y}$ , qui doit être égal à  $\mathbf{Z}^{*\top} \mathbf{M}_X \mathbf{u}$  sous  $H_1$  et doit donc avoir une espérance nulle, l'est vraiment.<sup>5</sup> Une forme quadratique de ce vecteur permet de construire une statistique de test qui obéit à une loi du  $\chi^2$ , mais on remarque que n'importe quelle statistique asymptotiquement équivalente à

$$\frac{1}{\sigma_1^2} \mathbf{u}^\top \mathbf{M}_X \mathbf{Z}^* (\mathbf{Z}^{*\top} \mathbf{M}_X \mathbf{Z}^*)^{-1} \mathbf{Z}^{*\top} \mathbf{M}_X \mathbf{u}$$

<sup>5</sup> Si  $\mathbf{X}$  ou  $\mathbf{Z}$  est composée de variables dépendantes retardées, alors nous sommes intéressés par l'espérance asymptotique de  $n^{-1/2} \mathbf{Z}^{*\top} \mathbf{M}_X \mathbf{y}$  plutôt que par l'espérance asymptotique de  $\mathbf{Z}^{*\top} \mathbf{M}_X \mathbf{y}$ .

le permet également. Mais cette statistique de test est bien sûr équivalente à un test en  $F$  ordinaire pour  $\gamma^* = \mathbf{0}$  dans la régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\boldsymbol{\gamma}^* + \mathbf{u}. \quad (11.24)$$

Ainsi il apparaît que, dans ce cas, le test d'enveloppement n'est rien de plus qu'un test en  $F$  ordinaire de  $H_1$  contre l'hypothèse alternative (11.24). Un tel test est facile à mettre en œuvre et sera exact sous les conditions habituelles.

Les mérites relatifs des tests à degré de liberté unique comme le test en  $J$  et des tests à degrés de liberté multiples comme le test d'enveloppement ont été largement développés dans la littérature; consulter Pesaran (1982) et l'article de synthèse de MacKinnon (1983), tout particulièrement les commentaires des nombreux intervenants. Le test en  $J$  et les tests équivalents seront plus puissants que les tests à degrés de liberté multiples lorsque les données ont été réellement générées par  $H_2$  mais peuvent être moins puissants lorsqu'elles sont générées par un tout autre modèle. Nous verrons les raisons de tout cela au cours du Chapitre 12, lorsque nous discuterons des éléments qui déterminent la puissance d'un test.

Dans la suite de cette section, nous discuterons de deux cas particuliers. Le premier concerne les modèles de régression dont les aléas sont autocorrélés. Même si un modèle de régression est linéaire à l'origine, sa transformation pour prendre en compte un processus AR(1) ou tout autre processus suivi par les aléas en fait un modèle non linéaire, ainsi que nous l'avons vu au cours du Chapitre 10. Supposons donc que les deux modèles concurrents soient

$$H_1: y_t = \mathbf{X}_t\boldsymbol{\beta} + u_{1t}, \quad u_{1t} = \rho_1 u_{1,t-1} + \varepsilon_{1t}, \quad \text{et}$$

$$H_2: y_t = \mathbf{Z}_t\boldsymbol{\gamma} + u_{2t}, \quad u_{2t} = \rho_2 u_{2,t-1} + \varepsilon_{2t}.$$

La manière la plus simple de procéder consiste à transformer ces modèles en des modèles de régression non linéaire tels que

$$H_1: y_t = \rho_1 y_{t-1} + (\mathbf{X}_t - \rho_1 \mathbf{X}_{t-1})\boldsymbol{\beta} + \varepsilon_{1t} \quad \text{et}$$

$$H_2: y_t = \rho_2 y_{t-1} + (\mathbf{Z}_t - \rho_2 \mathbf{Z}_{t-1})\boldsymbol{\gamma} + \varepsilon_{2t},$$

valables pour les observations allant de l'observation 2 à l'observation  $n$ . Alors on peut faire usage des tests en  $P$  ou en  $P_A$  pour tester  $H_1$  contre  $H_2$ , ou vice versa.

Remarquons que pour disposer des estimations  $(\tilde{\gamma}, \tilde{\rho}_2)$  nécessaires au test en  $P_A$  de  $H_1$  contre  $H_2$ , il faut exécuter la régression non linéaire

$$\hat{\rho}_1 y_{t-1} + (\mathbf{X}_t - \hat{\rho}_1 \mathbf{X}_{t-1})\hat{\boldsymbol{\beta}} = \rho_2 y_{t-1} + (\mathbf{Z}_t - \rho_2 \mathbf{Z}_{t-1})\boldsymbol{\gamma} + \varepsilon_{2t}. \quad (11.25)$$

Cette étape serait franchie grâce à un algorithme général d'estimation par NLS, puisque les algorithmes qui réalisent les procédures de Cochrane-Orcutt

ou de Hildreth-Lu utilisent  $\hat{\rho}_1 y_{t-1} + (\mathbf{X}_t - \hat{\rho}_1 \mathbf{X}_{t-1})\hat{\boldsymbol{\beta}}$  retardé une fois plutôt que  $y_{t-1}$  dans le membre de droite de (11.25). Bernanke, Bohn, et Reiss (1988) et McAleer, Pesaran, et Bera (1990) ont discuté des nombreuses procédures de test des modèles non emboîtés avec autocorrélation et les ont comparées en utilisant des simulations par la méthode de Monte Carlo.

Le second cas particulier qui nous intéresse concerne les modèles de régression estimés par variables instrumentales. Ericsson (1983) et Godfrey (1983) discutent des moyens nombreux et variés de manipuler de tels modèles. L'approche la plus simple, suggérée par MacKinnon, White, et Davidson (1983), consiste simplement à modifier les tests en  $J$  et en  $P$  de façon à les rendre adaptés à ce cas. La régression du test en  $P$  (11.15) devient

$$\mathbf{y} - \hat{\mathbf{x}} = \mathbf{P}_W \hat{\mathbf{X}} \mathbf{b} + a \mathbf{P}_W (\hat{\mathbf{z}} - \hat{\mathbf{x}}) + \text{résidus}, \quad (11.26)$$

où  $\hat{\mathbf{x}}$ ,  $\hat{\mathbf{X}}$ , et  $\hat{\mathbf{z}}$  sont désormais évalués avec les estimations IV  $\hat{\boldsymbol{\beta}}$  et  $\hat{\boldsymbol{\gamma}}$ . La manière la plus facile d'obtenir une statistique de test est simplement de régresser  $\mathbf{y} - \hat{\mathbf{x}}$  sur  $\hat{\mathbf{X}}$  et  $\hat{\mathbf{z}} - \hat{\mathbf{x}}$  par une procédure IV où  $\mathbf{W}$  est la matrice des instruments. La statistique pseudo- $t$  de l'estimation de  $a$  sera alors une statistique de test valable asymptotiquement, pourvu que  $\mathbf{W}$  soit l'ensemble des instruments avec lesquels on a estimé  $H_1$  par IV et que les conditions habituelles de régularité pour l'estimation par IV non linéaire soient satisfaites (consulter la Section 7.6).

Cela complète notre discussion sur les tests d'hypothèses non emboîtés pour les modèles de régression. À l'évidence, nous n'avons pas discuté de tous les aspects de ce problème. Les aspects dont nous n'avons pas parlé sont traités dans deux articles de MacKinnon, White, et Davidson (1983), qui adaptent les tests en  $J$  et en  $P$  aux modèles qui impliquent des transformations de la variable dépendante, et de Davidson et MacKinnon (1983b), qui adaptent ces mêmes tests aux modèles de régression non linéaire multivariée (voir Chapitre 9). Les études de MacKinnon (1983) et McAleer (1987) fournissent de nombreuses autres références.

## 11.4 TESTS BASÉS SUR DEUX ESTIMATIONS COMPARÉES

Dans la Section 7.9, nous avons introduit une classe de tests, que nous avons appelés tests de Durbin-Wu-Hausman, ou tests DWH, et qui peuvent être utilisés pour savoir si les estimations par moindres carrés sont convergentes lorsque certains régresseurs peuvent être corrélés aux termes d'erreur. Ces tests ont été développés par Durbin (1954), Wu (1973), et Hausman (1978). Il y a eu un important travail réalisé sur les tests DWH au cours des années récentes; voir l'article de synthèse de Ruud (1984). Dans cette section, nous montrons que les tests DWH peuvent se révéler utiles dans un grand nombre de circonstances non relatives à l'estimation IV, bien que l'on reste dans le contexte des modèles de régression.

L'idée de base des tests DWH est de construire un test sur un **vecteur de contraste**, c'est-à-dire la différence entre deux ensembles d'estimations, dont l'un sera convergent sous des conditions moins restrictives que l'autre. Supposons simplement que le modèle qu'il faut tester est

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (11.27)$$

où il y a  $n$  observations et  $k$  régresseurs. Dans ce contexte, le principe de test DWH suggère de comparer l'estimateur OLS

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (11.28)$$

avec un autre estimateur linéaire

$$\check{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{A} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{A} \mathbf{y}, \quad (11.29)$$

où  $\mathbf{A}$  est une matrice symétrique de dimension  $n \times n$  qui est supposée, pour simplifier, avoir un rang au moins égal à  $k$  (autrement, on ne pourrait pas estimer toutes les composantes de  $\check{\boldsymbol{\beta}}$ , et nous ne pourrions comparer que la partie estimée de  $\check{\boldsymbol{\beta}}$  au sous-vecteur correspondant de  $\hat{\boldsymbol{\beta}}$ ; voir la discussion sur les tests de spécification des dérivées qui suit). Dans le cas étudié au cours de la Section 7.9,  $\check{\boldsymbol{\beta}}$  est l'estimateur IV

$$\tilde{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y}.$$

Ainsi, dans ce cas la matrice  $\mathbf{A}$  correspond à  $\mathbf{P}_W$ , la matrice qui projette orthogonalement sur  $\mathcal{S}(\mathbf{W})$ , où  $\mathbf{W}$  est la matrice des instruments.

Si les données avaient été générées en réalité par le modèle (11.27), avec  $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ , les deux estimations (11.28) et (11.29) devraient avoir la même limite en probabilité. Pour s'en rendre compte, observons que

$$\text{plim}_{n \rightarrow \infty} \check{\boldsymbol{\beta}} = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{A} \mathbf{X} \right)^{-1} \left( \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{A} \mathbf{X} \right) \boldsymbol{\beta}_0 + \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{A} \mathbf{u} \right) \right),$$

qui est égal à  $\boldsymbol{\beta}_0$  à condition que  $\text{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{X}^\top \mathbf{A} \mathbf{u}) = \mathbf{0}$ . Ainsi, si  $\check{\boldsymbol{\beta}}$  et  $\hat{\boldsymbol{\beta}}$  diffèrent d'une quantité supérieure à ce que l'on peut raisonnablement attribuer à une variation aléatoire, on peut conclure que les données *n'ont pas* été générées par le modèle (11.27).

Pour un modèle de régression tel que (11.27), il est aisé de calculer un test DWH au moyen d'une régression artificielle. Nous avons vu des exemples similaires à la Section 7.9 et nous discuterons d'exemples plus éloignés plus tard. Cependant, il existe une autre façon de calculer des tests DWH, et cette autre façon peut être plus pratique dans certains cas. Pour un modèle quelconque qui n'est pas forcément un modèle de régression, supposons que  $\hat{\boldsymbol{\theta}}$  désigne un estimateur efficace des paramètres du modèle et que  $\check{\boldsymbol{\theta}}$  désigne



un estimateur moins efficace mais convergent sous des conditions moins restrictives que celles du modèle. Notons  $\mathbf{e}$  le vecteur de contraste entre  $\check{\boldsymbol{\theta}}$  et  $\hat{\boldsymbol{\theta}}$ . Alors nous avons vu que

$$n^{1/2}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + n^{1/2}\mathbf{e}, \quad (11.30)$$

où  $n^{1/2}\mathbf{e}$  est asymptotiquement non corrélé avec  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ . Ce résultat a été démontré pour les modèles estimés par maximum de vraisemblance, dans la Section 8.8; l'équivalent avec un échantillon fini pour les modèles de régression linéaire a été démontré en tant qu'élément de démonstration du Théorème de Gauss-Markov à la Section 5.5. Parce que les deux termes du membre de droite de (11.30) sont asymptotiquement non corrélés, la matrice de covariance asymptotique du membre de gauche correspond à la somme des deux matrices de covariance asymptotique de ces deux termes. Par conséquent, on obtient

$$\lim_{n \rightarrow \infty} \mathbf{V}(n^{1/2}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) = \lim_{n \rightarrow \infty} \mathbf{V}(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) + \lim_{n \rightarrow \infty} \mathbf{V}(n^{1/2}\mathbf{e}),$$

qui, en utilisant une notation simplifiée, nous permet de déduire la matrice de covariance asymptotique du vecteur de contraste:

$$\mathbf{V}^\infty(\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) = \mathbf{V}^\infty(\check{\boldsymbol{\theta}}) - \mathbf{V}^\infty(\hat{\boldsymbol{\theta}}). \quad (11.31)$$

Autrement dit, la matrice de covariance asymptotique de la différence entre  $\check{\boldsymbol{\theta}}$  et  $\hat{\boldsymbol{\theta}}$  égale la différence de leurs matrices de covariance asymptotique respectives. On doit ce résultat important à Hausman (1978).

On peut faire usage du résultat (11.31) pour construire des tests DWH de la forme

$$(\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top (\check{\mathbf{V}}(\check{\boldsymbol{\theta}}) - \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}))^{-1} (\check{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}), \quad (11.32)$$

où  $\check{\mathbf{V}}(\check{\boldsymbol{\theta}})$  et  $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$  désignent les estimations des matrices de covariance respectives de  $\check{\boldsymbol{\theta}}$  et  $\hat{\boldsymbol{\theta}}$ . La statistique de test (11.32) sera asymptotiquement distribuée suivant une loi du  $\chi^2(r)$  dont le nombre de degrés de liberté correspond au rang de  $\mathbf{V}^\infty(\check{\boldsymbol{\theta}}) - \mathbf{V}^\infty(\hat{\boldsymbol{\theta}})$ . Notons qu'il faudra remplacer l'inverse dans (11.32) par une inverse généralisée si, comme c'est souvent le cas, le rang de  $\mathbf{V}^\infty(\check{\boldsymbol{\theta}}) - \mathbf{V}^\infty(\hat{\boldsymbol{\theta}})$  est inférieur au nombre de paramètres de  $\boldsymbol{\theta}$ ; voir Hausman et Taylor (1982). Il peut survenir des difficultés d'ordre pratique avec (11.32) si  $\check{\mathbf{V}}(\check{\boldsymbol{\theta}}) - \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$  n'est pas semi-définie positive ou si le rang de  $\check{\mathbf{V}}(\check{\boldsymbol{\theta}}) - \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$  diffère de celui de  $\mathbf{V}^\infty(\check{\boldsymbol{\theta}}) - \mathbf{V}^\infty(\hat{\boldsymbol{\theta}})$ . C'est pour ces raisons que nous insistons sur l'approche basée sur les régressions artificielles.

Dans le cas de la régression linéaire (11.27), où les deux estimateurs sont (11.28) et (11.29), le test DWH est basé sur le vecteur de contraste

$$\check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{A} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{A} \mathbf{M}_X \mathbf{y}. \quad (11.33)$$

Cette expression ressemble justement à (7.59), avec  $\mathbf{A}$  en lieu et place de  $\mathbf{P}_W$ , et peut être dérivée exactement de la même manière. Le premier facteur

dans (11.33),  $(\mathbf{X}^\top \mathbf{A} \mathbf{X})^{-1}$ , est simplement une matrice de dimension  $k \times k$  de plein rang, qui sera sans influence sur les statistiques de test que l'on pourrait calculer. Par conséquent, ce que nous désirons réellement tester, c'est si le vecteur

$$n^{-1/2} \mathbf{X}^\top \mathbf{A} \mathbf{M}_X \mathbf{y} \quad (11.34)$$

a une espérance nulle, asymptotiquement. Ce vecteur est composé de  $k$  éléments, mais même si  $\mathbf{A} \mathbf{X}$  est de plein rang, tous les éléments ne sont pas des variables aléatoires, parce que  $\mathbf{M}_X$  peut annuler certaines colonnes de  $\mathbf{A} \mathbf{X}$ . Supposons que  $k^*$  est le nombre de colonnes de  $\mathbf{A} \mathbf{X}$  linéairement indépendantes qui ne sont pas annulées par  $\mathbf{M}_X$ . Alors le test de (11.34) est équivalent au test de nullité asymptotique de l'espérance de

$$n^{-1/2} \mathbf{X}^{*\top} \mathbf{A} \mathbf{M}_X \mathbf{y} \quad (11.35)$$

où  $\mathbf{X}^*$  est la matrice des  $k^*$  colonnes de  $\mathbf{X}$  telles qu'aucune colonne de  $\mathbf{A} \mathbf{X}^*$  n'est annulée par  $\mathbf{M}_X$ .

Considérons à présent la régression artificielle

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{A} \mathbf{X}^* \boldsymbol{\delta} + \text{résidus}. \quad (11.36)$$

On montre aisément grâce au Théorème FWL que l'estimation OLS de  $\boldsymbol{\delta}$  est

$$\hat{\boldsymbol{\delta}} = (\mathbf{X}^{*\top} \mathbf{A} \mathbf{M}_X \mathbf{A} \mathbf{X}^*)^{-1} \mathbf{X}^{*\top} \mathbf{A} \mathbf{M}_X \mathbf{y},$$

et il est évident que, en général,  $\text{plim}(\hat{\boldsymbol{\delta}}) = \mathbf{0}$  si et seulement si (11.35) est exacte. Le  $F$  de Fisher ordinaire pour  $\boldsymbol{\delta} = \mathbf{0}$  dans (11.36) est

$$\frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{M}_X \mathbf{A} \mathbf{X}^*} \mathbf{y} / k^*}{\mathbf{y}^\top \mathbf{M}_{\mathbf{X}, \mathbf{M}_X \mathbf{A} \mathbf{X}^*} \mathbf{y} / (n - k - k^*)}, \quad (11.37)$$

où  $\mathbf{P}_{\mathbf{M}_X \mathbf{A} \mathbf{X}^*}$  est la matrice qui projette orthogonalement sur  $\mathcal{S}(\mathbf{M}_X \mathbf{A} \mathbf{X}^*)$ , et  $\mathbf{M}_{\mathbf{X}, \mathbf{M}_X \mathbf{A} \mathbf{X}^*}$  est la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}, \mathbf{M}_X \mathbf{A} \mathbf{X}^*)$ . Si (11.27) a réellement généré les données, la statistique (11.37) sera certainement valable asymptotiquement, puisque le dénominateur sera une estimation convergente de  $\sigma^2$ . Elle sera exactement distribuée suivant une  $F(k^*, n - k - k^*)$  avec des échantillons finis si les aléas dans (11.27) sont normalement distribuées et si  $\mathbf{X}$  et  $\mathbf{A}$  peuvent être considérées comme fixes. La régression (11.36) et l'expression (11.37) sont pour l'essentiel les mêmes que la régression (7.62) et l'expression (7.64), respectivement; ces dernières ne sont que des cas particuliers des premiers.

Le type de test DWH le plus fréquent est celui que nous avons examiné à la Section 7.9, qui permet de savoir si des estimations par moindres carrés sont convergentes lorsque quelques régresseurs peuvent être corrélés aux aléas.

Cependant, il existe de nombreuses autres éventualités. Par exemple,  $\check{\beta}$  pourrait être l'estimateur OLS de  $\beta$  pour le modèle

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{u}, \quad (11.38)$$

où  $\mathbf{Z}$  est une matrice de dimension  $n \times l$  composée de régresseurs qui ne sont pas dans l'espace engendré par les colonnes de  $\mathbf{X}$ . L'usage du Théorème FWL nous montre que

$$\check{\beta} = (\mathbf{X}^\top \mathbf{M}_Z \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M}_Z \mathbf{y},$$

expression dans laquelle  $\mathbf{M}_Z$  joue le rôle de  $\mathbf{A}$ . Cette forme du test DWH permet ainsi de savoir si les estimations  $\check{\beta}$ , lorsque  $\mathbf{Z}$  est incluse dans le modèle, diffèrent de façon significative des estimations  $\hat{\beta}$  lorsque  $\mathbf{Z}$  ne l'est pas. Ceci est un exemple simplifié du cas examiné par Holly (1982), dans un contexte beaucoup plus général. Il apparaît donc que cette version du test DWH est équivalente à un test en  $F$  ordinaire pour  $\gamma = \mathbf{0}$ , à condition que  $k \geq l$  et qu'une certaine matrice soit de plein rang, mais ne l'est pas sinon. On peut voir cela à partir de la régression (11.36), qui est dans ce cas

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{M}_Z \mathbf{X}\delta + \text{résidus} \quad (11.39)$$

$$= \mathbf{X}(\beta + \delta) - \mathbf{P}_Z \mathbf{X}\delta + \text{résidus}. \quad (11.40)$$

Il est évident à partir de (11.40) que chaque fois que la matrice  $\mathbf{Z}^\top \mathbf{X}$  sera de rang  $l$ , la régression (11.39) aura exactement le même pouvoir explicatif que la régression (11.38), puisque  $\mathbf{X}$  et  $\mathbf{P}_Z \mathbf{X} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}$  engendreront conjointement le même sous-espace que  $\mathbf{X}$  et  $\mathbf{Z}$ . Le test en  $F$  pour  $\delta = \mathbf{0}$  dans (11.39) sera ainsi identique au test en  $F$  pour  $\gamma = \mathbf{0}$  dans (11.38), ce qui est le résultat obtenu par Holly dans le cas très particulier de la régression linéaire. Une condition nécessaire, mais pas suffisante, pour que  $\mathbf{Z}^\top \mathbf{X}$  soit de rang  $l$ , est que  $k \geq l$ . Pour de plus amples détails sur la relation entre les tests DWH et les tests d'hypothèses classiques, consulter Holly et Monfort (1986) et Davidson et MacKinnon (1989).

Il y a une relation intéressante entre la variante d'"exogénéité" du test DWH et la variante "variables omises". Dans la première,  $\mathbf{A} = \mathbf{P}_W$  et  $\mathbf{P}_W \mathbf{X}^*$  est composée de toutes les colonnes de  $\mathbf{P}_W \mathbf{X}$  qui n'appartiennent pas à l'espace engendré par les colonnes de  $\mathbf{X}$ . Ainsi la régression du test est

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{P}_W \mathbf{X}^* \delta + \text{résidus}. \quad (11.41)$$

Dans cette dernière,  $\mathbf{M}_Z \mathbf{X}^* = \mathbf{M}_Z \mathbf{X}$ , pourvu que la matrice  $[\mathbf{X} \ \mathbf{Z}]$  soit de plein rang. Supposons désormais que l'on développe  $\mathbf{Z}$  de manière à la rendre égale à  $\mathbf{W}$ , ce qui signifie qu'elle comprend au moins autant de variables que  $\mathbf{X}$ , et parmi elles certaines variables qui n'appartiennent pas à l'espace engendré par  $\mathbf{X}$ . Evidemment,  $\mathbf{X}^*$  sera alors composée de ces colonnes de  $\mathbf{X}$

qui n'appartiennent pas à l'espace engendré par  $\mathbf{W}$ , et la régression de test sera

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}_W\mathbf{X}^*\boldsymbol{\delta} + \text{résidus.} \quad (11.42)$$

Parce que les matrices  $[\mathbf{X} \ \mathbf{P}_W\mathbf{X}]$  et  $[\mathbf{X} \ \mathbf{M}_W\mathbf{X}]$  engendrent le même sous-espace, les régressions (11.41) et (11.42) auront exactement le même pouvoir explicatif. Cela signifie que le test que l'on *interprète* comme un test de convergence en présence d'une endogénéité éventuelle et le test que l'on *interprète* comme un test de convergence des estimations des paramètres lorsque certaines variables ont été omises sont en réalité exactement les mêmes. Ruud (1984) détaille davantage la discussion.

L'ultime exemple des tests DWH dont nous allons discuter est le **test de spécification des différences** qui a été proposé par Plosser, Schwert, et White (1982). L'idée de base de ce test est de construire un test de spécification fondé sur la comparaison des estimations en niveau et des estimations aux premières différences. Notre traitement fait suite à celui de Davidson, Godfrey, et MacKinnon (1985), qui montre la manière de calculer le test à l'aide d'une régression artificielle.

Comme d'habitude, l'estimation OLS en niveau est  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ . L'estimation OLS utilisant les données différenciées une fois est

$$\check{\boldsymbol{\beta}} = (\dot{\mathbf{X}}^\top \dot{\mathbf{X}})^{-1} \dot{\mathbf{X}}^\top \dot{\mathbf{y}},$$

où  $\dot{\mathbf{y}}$  et  $\dot{\mathbf{X}}$  désignent le vecteur et la matrice dont les lignes types respectives sont  $\dot{y}_t = y_t - y_{t-1}$  et  $\dot{\mathbf{X}}_t = \mathbf{X}_t - \mathbf{X}_{t-1}$ . Pour l'instant nous ne nous soucions pas du fait que si  $\mathbf{X}$  comprend une constante,  $\dot{\mathbf{X}}$  possédera une colonne composée de zéros. Nous ignorons également le fait que l'on ne peut pas calculer  $\dot{\mathbf{X}}_1$  et  $\dot{\mathbf{y}}_1$  sans recourir à des hypothèses arbitraires si  $\mathbf{X}_0$  et  $\mathbf{y}_0$  ne sont pas disponibles.

Le résultat crucial qui rend possible le calcul du test de différentiation au moyen d'une régression artificielle, est que, si  $\ddot{\mathbf{X}}$  désigne la matrice dont la ligne type est  $\mathbf{X}_{t+1} - 2\mathbf{X}_t + \mathbf{X}_{t-1}$  (c'est-à-dire la matrice des différences secondes de  $\mathbf{X}$ , avancées d'une période), alors

$$\check{\boldsymbol{\beta}} \stackrel{a}{=} (-\ddot{\mathbf{X}}^\top \mathbf{X})^{-1} (-\ddot{\mathbf{X}}^\top \mathbf{y}) = (\ddot{\mathbf{X}}^\top \mathbf{X})^{-1} \ddot{\mathbf{X}}^\top \mathbf{y}. \quad (11.43)$$

Pour démontrer cela, considérons les éléments types des matrices qui apparaissent dans (11.43). Supposons que  $\mathbf{r}$  désigne n'importe quelle colonne de  $\mathbf{X}$  et  $\mathbf{s}$  désigne la même colonne ou n'importe quelle autre colonne de  $\mathbf{X}$ , ou éventuellement  $\mathbf{y}$ . Par conséquent n'importe quel élément de  $\ddot{\mathbf{X}}^\top \mathbf{X}$ , ou de  $\ddot{\mathbf{X}}^\top \dot{\mathbf{y}}$ , peut être écrit  $\dot{\mathbf{r}}^\top \dot{\mathbf{s}}$ , alors que n'importe quel élément de  $\ddot{\mathbf{X}}^\top \mathbf{X}$ , ou de  $\ddot{\mathbf{X}}^\top \mathbf{y}$ , peut être écrit  $\ddot{\mathbf{r}}^\top \mathbf{s}$ . Nous voulons montrer que  $\dot{\mathbf{r}}^\top \dot{\mathbf{s}} \stackrel{a}{=} -\ddot{\mathbf{r}}^\top \mathbf{s}$ . Par définition

$$\begin{aligned} \dot{\mathbf{r}}^\top \dot{\mathbf{s}} &= \sum_{t=1}^n (r_t - r_{t-1})(s_t - s_{t-1}) \\ &= \sum_{t=1}^n (r_t s_t + r_{t-1} s_{t-1} - r_t s_{t-1} - r_{t-1} s_t). \end{aligned} \quad (11.44)$$

De façon similaire

$$\begin{aligned} -\ddot{\mathbf{r}}^\top \mathbf{s} &= -\sum_{t=1}^n (r_{t+1} - 2r_t + r_{t-1}) s_t \\ &= \sum_{t=1}^n (2r_t s_t - r_{t+1} s_t - r_{t-1} s_t). \end{aligned} \quad (11.45)$$

Soustraire (11.45) à (11.44) entraîne

$$r_0 s_0 - r_n s_n - r_1 s_0 + r_{n+1} s_n.$$

Cette expression est évidemment  $O(1)$ , alors que des quantités telles que  $\dot{\mathbf{X}}^\top \dot{\mathbf{X}}$  et  $\ddot{\mathbf{X}}^\top \mathbf{X}$  sont  $O(n)$ . Toute différence entre  $\dot{\mathbf{r}}^\top \dot{\mathbf{s}}$  et  $-\ddot{\mathbf{r}}^\top \mathbf{s}$  doit par conséquent être asymptotiquement négligeable, ce qui démontre le résultat (11.43).<sup>6</sup>

Grâce à ce résultat et au fait que  $\mathbf{y} = \mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}_X \mathbf{y}$ , nous apercevons que

$$\begin{aligned} \check{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} &\stackrel{a}{=} (\ddot{\mathbf{X}}^\top \mathbf{X})^{-1} \ddot{\mathbf{X}}^\top \mathbf{y} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= (\ddot{\mathbf{X}}^\top \mathbf{X})^{-1} \ddot{\mathbf{X}}^\top (\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}_X \mathbf{y}) - \hat{\boldsymbol{\beta}} \\ &= (\ddot{\mathbf{X}}^\top \mathbf{X})^{-1} \ddot{\mathbf{X}}^\top \mathbf{M}_X \mathbf{y}. \end{aligned}$$

Ainsi le test de spécification des différences est vraiment un test de l'hypothèse que le vecteur  $n^{-1/2} \ddot{\mathbf{X}}^\top \mathbf{M}_X \mathbf{y}$  a une espérance nulle asymptotiquement. Par un argument similaire à celui qui conduit à la régression artificielle (11.36), il est aisé de montrer que cette hypothèse peut être testée grâce à un test en  $F$  ordinaire pour  $\boldsymbol{\delta} = \mathbf{0}$  dans la régression artificielle

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \ddot{\mathbf{X}}\boldsymbol{\delta} + \text{résidus}. \quad (11.46)$$

De plus, d'après la définition de  $\ddot{\mathbf{X}}$  nous voyons que  $\mathcal{S}(\mathbf{X}, \ddot{\mathbf{X}}) = \mathcal{S}(\mathbf{X}, \mathbf{C})$ , où  $\mathbf{C}$  est une matrice dont la ligne type est  $\mathbf{X}_{t-1} + \mathbf{X}_{t+1}$ . Ainsi le test pour  $\boldsymbol{\delta} = \mathbf{0}$  dans (11.46) sera numériquement identique au test pour  $\boldsymbol{\eta} = \mathbf{0}$  dans

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{C}\boldsymbol{\eta} + \text{résidus}. \quad (11.47)$$

La régression (11.47) permet de voir la manière de traiter la constante et tout régresseur appartenant à  $\mathbf{X}$  qui, après avoir pris les différences premières, empêche la matrice  $\ddot{\mathbf{X}}$  d'être de plein rang. Si un tel régresseur est inclus dans la fonction de régression, la matrice  $[\mathbf{X} \ \mathbf{C}]$  ne sera pas de plein rang. Il faut

<sup>6</sup> On peut également démontrer ce résultat à l'aide d'une **matrice de différences**, disons  $\mathbf{D}$ , telle que  $\dot{\mathbf{X}} = \mathbf{D}\mathbf{X}$  et  $\ddot{\mathbf{X}}_{-1} = \mathbf{D}^2 \mathbf{X}$ . Une telle démonstration serait plus concise mais sans doute moins facile à saisir.

donc éliminer de  $\mathbf{C}$  toutes les colonnes qui empêchent  $[\mathbf{X} \ \mathbf{C}]$  d'être de plein rang. Le nombre de degrés de liberté pour la statistique de test correspondra alors au nombre de colonnes restantes de  $\mathbf{C}$ .

La régression (11.47) montre également que le test de spécification de la différentiation est en réalité un test curieux. Les régresseurs supplémentaires dans  $\mathbf{C}$  sont les sommes des valeurs avancées et retardées des régresseurs originels. Bien qu'il soit aisé de justifier le test de l'inclusion éventuelle des valeurs retardées de  $\mathbf{X}$  dans le modèle de régression, il est plus délicat de justifier le test de l'inclusion des valeurs avancées de  $\mathbf{X}$ . Dans de nombreux cas, on ne s'attend pas à ce que l'ensemble des informations qui conditionne  $\mathbf{y}$  contienne des valeurs avancées de  $\mathbf{X}$ . Assurément le test sera non pertinent si  $\mathbf{X}$  peut dépendre des valeurs retardées de  $\mathbf{u}$ , puisque dans ce cas  $u_t$  peut être corrélé à  $\mathbf{X}_{t+1}$ .

Il existe de nombreuses autres applications du test DWH aux modèles de régression linéaire et non linéaire. Consulter Boothe et MacKinnon (1986), Breusch et Godfrey (1986), Godfrey (1988), et Ruud (1984). Nous avons discuté des tests de la différence entre des estimations IV et moindres carrés des modèles de régression non linéaire dans la Section 7.9, et la majeure partie des arguments est valable pour les autres applications du test DWH aux modèles de régression non linéaire.

On affirme souvent que les tests DWH peuvent être utilisés à profit lorsque l'hypothèse nulle *n'est pas* que les données ont été générées par (11.27) mais simplement que les estimations OLS  $\hat{\beta}$  de (11.27) sont convergentes. Bien que cela soit vrai jusqu'à un certain point, il existe une difficulté réelle lorsque l'on essaye d'utiliser ces tests dans cette optique. Ainsi que nous l'avons vu, les tests DWH ne testent pas directement l'hypothèse selon laquelle les paramètres sont estimés de façon convergente. Au lieu de cela, ils testent la nullité de certaines combinaisons linéaires des paramètres sur certaines variables omises, parce que si c'était effectivement le cas, cela impliquerait que les paramètres de l'hypothèse nulle sont estimés de façon convergente. Par conséquent, il y a des situations où tous les paramètres seront estimés de façon convergente et malgré cela les tests DWH rejettent presque invariablement l'hypothèse nulle.

Pour apercevoir la manière dont cela peut survenir, considérons le cas très simple suivant. Supposons que le modèle contraint soit

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \quad (11.48)$$

et que le modèle non contraint soit

$$\mathbf{y} = \mathbf{X}\beta + \gamma\mathbf{z} + \mathbf{u}, \quad (11.49)$$

avec  $\mathbf{X}$  une matrice aléatoire de dimension  $n \times k$ ,  $\mathbf{z}$  et  $\mathbf{u}$  deux vecteurs aléatoires à  $n$  éléments distribués de telle sorte que  $\text{plim}(n^{-1}\mathbf{X}^\top\mathbf{z}) = \mathbf{0}$  et

$\text{plim}(n^{-1}\mathbf{X}^\top \mathbf{u}) = \mathbf{0}$ . Il est clair que l'estimation OLS de (11.48) entraînera des estimations convergentes de  $\beta$  même si le DGP est (11.49) où  $\gamma \neq 0$ . Considérons à présent le test DWH qui pourrait être basé sur la régression

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{z}(\mathbf{z}^\top \mathbf{z})^{-1} \mathbf{z}^\top \mathbf{x}^* \delta + \text{résidus}, \quad (11.50)$$

où  $\mathbf{x}^*$  est l'une des colonnes de  $\mathbf{X}$ . A moins que  $\mathbf{z}^\top \mathbf{x}^*$  ne soit numériquement nul, auquel cas on ne peut pas calculer le test, un test en  $t$  pour  $\delta = 0$  sera identique numériquement à un test en  $t$  pour  $\gamma = 0$  dans (11.49). Ainsi, si  $\gamma \neq 0$  et si l'échantillon est suffisamment large, le test DWH rejettera l'hypothèse nulle avec une probabilité égale à l'unité, même dans le cas où  $\hat{\beta}$  est en réalité convergent. La raison de ce problème embarrassant en apparence est que nous avons calculé un test DWH avec un échantillon fini qu'il aurait été impossible de calculer asymptotiquement, parce que le régresseur  $\mathbf{z}(\mathbf{z}^\top \mathbf{z})^{-1} \mathbf{z}^\top \mathbf{x}^*$  serait alors une colonne de zéros. Malheureusement, on peut souvent calculer ce test. Dans de telles circonstances, il est clair que les résultats des tests DWH, avec des échantillons finis, peuvent être mal interprétés.

## 11.5 TESTS D'HÉTÉROSCÉDASTICITÉ

Tous les tests basés sur la régression de Gauss-Newton dont nous avons discuté jusqu'à présent sont conçus pour tester des aspects variés de la spécification des fonctions de régression. Cependant, des variantes de la GNR peuvent également être utilisées pour tester certains aspects de la spécification des aléas, en particulier l'hypothèse d'une variance constante. Dans cette section, nous allons montrer comment on peut dériver certains tests d'hétéroscédasticité très répandus, comme des applications de la GNR. Nous discuterons des tests d'hétéroscédasticité supplémentaires au cours du Chapitre 16.

Un modèle d'hétéroscédasticité plausible est

$$E(u_t^2) = h(\alpha + \mathbf{Z}_t \gamma), \quad (11.51)$$

où  $h(\cdot)$  est une fonction éventuellement non linéaire qui doit produire des valeurs positives,  $\mathbf{Z}_t$  est un vecteur dont les  $q$  composantes sont des observations sur des variables exogènes ou prédéterminées,  $\alpha$  est un scalaire,  $\gamma$  un vecteur de  $q$  paramètres. L'équation (11.51) indique que l'espérance de l'aléa  $u_t$  au carré est  $h(\alpha + \mathbf{Z}_t \gamma)$ . Comme nous l'avons vu à la Section 9.2, la fonction  $h(\cdot)$  est appelée **fonction scédastique**. Si tous les éléments du vecteur  $\gamma$  sont nuls,  $h(\alpha + \mathbf{Z}_t \gamma)$  se réduit à  $h(\alpha)$ , qui est simplement une constante. On peut imaginer que cette constante est  $\sigma^2$ . Ainsi nous pourrions tester l'hypothèse nulle d'homoscédasticité contre l'hypothèse alternative d'hétéroscédasticité (11.51) en testant la contrainte  $\gamma = \mathbf{0}$ .

Définissons désormais  $e_t$  comme la différence entre  $u_t^2$  et son espérance. Cela nous permet d'écrire une équation pour  $u_t^2$ :

$$u_t^2 = h(\alpha + \mathbf{Z}_t\gamma) + e_t. \quad (11.52)$$

L'équation (11.52) est un modèle de régression. Bien que l'on ne puisse pas s'attendre à ce que l'aléa  $e_t$  ait un comportement aussi régulier que la plupart des aléas des modèles de régression, puisque la distribution de  $u_t^2$  sera en général inclinée à droite, il doit avoir une espérance nulle par définition, et nous supposons qu'il a une variance finie et constante. Cette hypothèse serait probablement excessivement forte si  $\gamma$  était non nul (on pourra la relâcher en faisant usage des techniques discutées dans la prochaine section). Sous l'hypothèse nulle que  $\gamma = \mathbf{0}$  cependant, il ne paraît pas déraisonnable de supposer que la variance de  $e_t$  est constante.

Supposons pour débiter que l'on observe réellement  $u_t$ . Alors on peut sûrement estimer (11.52) à la manière habituelle par NLS. Sous l'hypothèse nulle que  $\gamma = \mathbf{0}$ , l'estimation NLS de  $\alpha$  est la valeur  $\tilde{\alpha}$  qui vérifie l'équation

$$h(\tilde{\alpha}) = \frac{1}{n} \sum_{t=1}^n u_t^2 \equiv \tilde{\sigma}^2.$$

Ainsi il suffit d'estimer la moyenne d'échantillonnage des  $u_t^2$ ,  $\tilde{\sigma}^2$ . On pourrait ensuite tester l'hypothèse que  $\gamma = \mathbf{0}$  au moyen d'une régression de Gauss-Newton. Cette GNR serait

$$u_t^2 - \tilde{\sigma}^2 = h'(\tilde{\alpha})a + h'(\tilde{\alpha})\mathbf{Z}_t\mathbf{c} + \text{résidu}, \quad (11.53)$$

où  $h'(\tilde{\alpha})$  est la dérivée de  $h(\cdot)$  par rapport à son unique argument, évaluée en  $\alpha = \tilde{\alpha}$  et  $\gamma = \mathbf{0}$ . Puisque  $h'(\tilde{\alpha})$  est une constante, (11.53) se simplifie et devient

$$\mathbf{v} - \iota\tilde{\sigma}^2 = \iota a + \mathbf{Z}\mathbf{c} + \text{résidus}, \quad (11.54)$$

où  $\mathbf{v}$  est un vecteur dont les  $n$  éléments sont les  $u_t^2$ ,  $\iota$  est un vecteur dont chaque composante égale 1, et  $\mathbf{Z}$  est une matrice de dimension  $n \times q$  dont la ligne type est  $\mathbf{Z}_t$ . Puisque ni la fonction  $h(\cdot)$  ni ses dérivées n'apparaissent dans (11.54), un test basé sur cette régression artificielle ne dépendra pas de la forme fonctionnelle de  $h(\cdot)$ . La raison est que tous les modèles de la forme (11.52) sont des alternatives localement équivalentes. Nous avons vu un exemple plus tôt à la Section 10.8; consulter Godfrey (1981) et Godfrey et Wickens (1982).

Comme d'habitude, la statistique de test de la GNR pour  $\gamma = \mathbf{0}$  est soit un test en  $F$  pour  $\mathbf{c} = \mathbf{0}$  dans (11.54) soit un  $nR^2$  de cette régression. Puisque  $\iota$  apparaît dans les deux membres de (11.54), la régression peut être encore simplifiée pour donner

$$\mathbf{v} = \iota a^* + \mathbf{Z}\mathbf{c} + \text{résidus}. \quad (11.55)$$



Le  $R^2$  centré de (11.55) sera identique à la fois au  $R^2$  centré et au  $R^2$  non centré de (11.54), qui sont les mêmes quantités parce que la régressande de (11.54) est d'espérance nulle par construction. La statistique  $F$  pour  $\mathbf{c} = \mathbf{0}$ , qui est rapporté par presque tous les progiciels de régression, sera bien évidemment identique pour les deux régressions.

Dans la pratique bien sûr, les aléas  $u_t$  apparaissent dans un modèle de régression comme  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}$ , et nous n'avons pas l'occasion de les observer véritablement. Toutefois, comme nous observons  $\mathbf{y}$  et tous les régresseurs qui nous semblent expliquer  $\mathbf{y}$ , on peut obtenir facilement les résidus des estimations par moindres carrés  $\hat{\mathbf{u}}$ . Le modèle qu'il faut estimer peut être linéaire ou non; la forme exacte est sans importance. Ainsi que nous l'avons vu à la Section 5.6, la convergence des estimations des paramètres par NLS implique que  $\hat{\mathbf{u}} \stackrel{a}{=} \mathbf{u}$ . Par conséquent, la régression

$$\hat{\mathbf{v}} = \boldsymbol{\iota}a^* + \mathbf{Z}\mathbf{c} + \text{résidus}, \quad (11.56)$$

où  $\hat{u}_t^2$  est l'élément type de  $\hat{\mathbf{v}}$ , générera des statistiques de test qui ont les mêmes propriétés asymptotiques que les statistiques de test générées par (11.55). Comme auparavant, un test en  $F$  ordinaire pour  $\mathbf{c} = \mathbf{0}$  sera asymptotiquement valable, autant que  $n$  fois le  $R^2$  centré.

Il peut paraître étonnant que l'on puisse remplacer  $\mathbf{v}$  par  $\hat{\mathbf{v}}$  sans rien faire pour tenir compte du fait que  $\boldsymbol{\beta}$  doit être estimé pour obtenir  $\hat{\mathbf{u}}$ , car lorsque nous utilisons une GNR pour un test de spécification de la fonction de régression, il nous faut justement en tenir compte. L'explication de cette différence devrait apparaître clairement à partir des deux exemples suivants. Premièrement, considérons les modèles de régression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{et} \quad (11.57)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{z} + \mathbf{u}. \quad (11.58)$$

Pour tester (11.57) contre (11.58), nous utiliserions normalement un  $t$  de Student, dont le numérateur serait

$$\mathbf{z}^\top \hat{\mathbf{u}} = \mathbf{z}^\top \mathbf{M}_X \mathbf{u} = \mathbf{z}^\top \mathbf{u} - \mathbf{z}^\top \mathbf{P}_X \mathbf{u}.$$

Puisqu'à la fois  $\mathbf{z}^\top \mathbf{u}$  et  $\mathbf{z}^\top \mathbf{P}_X \mathbf{u}$  sont  $O(n^{1/2})$ , il serait clairement erroné de traiter  $\mathbf{z}^\top \hat{\mathbf{u}}$  comme une quantité asymptotiquement équivalente à  $\mathbf{z}^\top \mathbf{u}$ . C'est la raison pour laquelle on peut calculer une statistique de test asymptotiquement valable en régressant  $\hat{\mathbf{u}}$  sur  $\mathbf{X}$  et  $\mathbf{z}$  mais pas en régressant  $\hat{\mathbf{u}}$  sur  $\mathbf{z}$  seulement.

Supposons maintenant que l'on veuille savoir si les aléas au carré de (11.57) sont corrélés avec  $\mathbf{z}$ . Souvenons-nous que  $\mathbf{v}$  est le vecteur des aléas au carré et  $\hat{\mathbf{v}}$  est le vecteur des résidus au carré. Si nous utilisons  $\hat{\mathbf{v}}$  comme un représentant de  $\mathbf{v}$  et le régressons sur la constante et sur  $\mathbf{z}$ , comme dans (11.56), le numérateur de ce  $t$  de Student est

$$\begin{aligned} \mathbf{z}^\top \mathbf{M}_\iota \hat{\mathbf{v}} &= \mathbf{z}^\top \mathbf{M}_\iota \mathbf{v} - 2\mathbf{z}^\top \mathbf{M}_\iota ((\mathbf{P}_X \mathbf{u}) * \mathbf{u}) + \mathbf{z}^\top \mathbf{M}_\iota ((\mathbf{P}_X \mathbf{u}) * (\mathbf{P}_X \mathbf{u})) \\ &= \mathbf{z}^\top \mathbf{M}_\iota \mathbf{v} + \mathbf{z}^\top \mathbf{M}_\iota ((\mathbf{P}_X \mathbf{u}) * (\mathbf{P}_X \mathbf{u} - 2\mathbf{u})), \end{aligned} \quad (11.59)$$

où  $\mathbf{M}_L$  est la matrice qui calcule les écarts à la moyenne, et  $*$  désigne le produit direct de deux vecteurs. Il est aisé de voir que le premier terme de la seconde ligne de (11.59) est  $O(n^{1/2})$ ; c'est simplement la somme de  $n$  termes, dont chacun est d'espérance nulle à cause de la présence de  $\mathbf{M}_L$ . Le second terme, par contre, est  $O(1)$ , ce qui signifie qu'il est asymptotiquement négligeable par rapport au premier. Ainsi  $\mathbf{z}^\top \mathbf{M}_L \hat{\mathbf{v}}$  est asymptotiquement équivalent à  $\mathbf{z}^\top \mathbf{M}_L \mathbf{v}$ , et l'on peut ignorer la distinction entre  $\mathbf{v}$  et  $\hat{\mathbf{v}}$  lorsque l'on calcule des tests d'hétéroscédasticité.

Une autre façon de considérer le problème est de se rappeler que, comme nous l'avons vu à la Section 8.10 lorsque nous avons discuté des modèles de régression non linéaire dans le contexte de l'estimation par maximum de vraisemblance, la matrice de covariance des estimations des paramètres d'un tel modèle est bloc-diagonale entre les paramètres de la fonction de régression (dans ce cas  $\beta$ ) et les paramètres de la fonction scédastique (dans ce cas  $\alpha$  et  $\gamma$ ). Cette propriété de bloc-diagonalité implique que l'on peut traiter les premiers paramètres comme connus dans le but de tester les seconds, et vice versa, même si on les estime en réalité.

Bien que la famille des tests que nous avons esquissée semble être une application naturelle de la régression de Gauss-Newton, ce n'est pas de cette façon qu'elle a été développée dans la littérature économétrique. Godfrey (1978c) et Breusch et Pagan (1979) ont proposé des statistiques de test qui, bien que fondées sur une légère modification de la régression artificielle (11.56), n'étaient pas les mêmes que celles que nous suggérons ici. Ces auteurs supposèrent explicitement que les aléas  $u_t$  étaient normalement distribués. Cela leur permit de dériver leurs tests comme des tests du multiplicateur de Lagrange en utilisant la théorie du maximum de vraisemblance, et ils obtinrent des statistiques de test quelque peu différentes qui restent valables même asymptotiquement, uniquement sous l'hypothèse de normalité. Koenker (1981) fit remarquer cette faiblesse des tests Godfrey/Breusch-Pagan et suggéra le test du  $nR^2$  basé sur la régression (11.56) comme alternative. Le test en  $F$  pour  $\mathbf{c} = \mathbf{0}$  basé sur la même régression est aussi valable asymptotiquement, et présente de nombreux attraits avec des échantillons finis. Malheureusement, les tests en  $F$  et  $nR^2$  peuvent souvent être moins puissants que les tests LM basés sur l'hypothèse de normalité. Honda (1988) a récemment montré la façon d'obtenir des versions modifiées de ces derniers et qui possèdent de meilleures propriétés avec des échantillons finis. Voir la Section 16.5 et Godfrey (1988, Section 4.5) pour une discussion plus complète de tous ces tests.

Au lieu de (11.51), on pourrait débiter avec le modèle plus général

$$E|u_t|^p = h(\alpha + \mathbf{Z}_t \gamma).$$

Glejser (1969) considéra le cas  $p = 1$  et proposa un test basé sur une régression artificielle similaire à (11.56) mais où la régressande est égale aux valeurs absolues des résidus. Dans l'article de Newey et Powell (1987), il est montré

que le test de Glejser peut gagner considérablement en puissance par rapport au test habituel, basé sur les carrés des résidus, dans le cas où les aléas ont des queues de distribution plus épaisses que celles de la distribution normale. Cela suggère qu'il peut être souvent très sage d'employer les deux types de tests.

## 11.6 UNE GNR ROBUSTE À L'HÉTÉROSCÉDASTICITÉ

Dans de nombreux cas, nous savons, ou du moins nous supposons, que les aléas associés à un modèle de régression manifestent de l'hétéroscédasticité, mais nous ne connaissons pas du tout la forme qu'elle prend. En particulier lorsque l'on travaille sur des données en coupe transversale, on présume que les aléas sont probablement hétéroscédastiques. Cela devrait nous mettre mal à l'aise sur l'usage des tests basés sur la régression de Gauss-Newton, ou sur l'usage de n'importe quel autre test dont nous avons discuté jusqu'ici puisqu'ils ne sont valables que sous l'hypothèse d'homoscédasticité. En réalité, il se trouve qu'il est assez simple de dériver une régression artificielle que l'on peut utiliser comme une GNR et qui produit des inférences asymptotiquement valables même en présence d'hétéroscédasticité dont la forme est inconnue. Dans cette section, nous discuterons de cette procédure, brièvement. Dans le Chapitre 16, nous offrirons un traitement plus complet sur ce sujet et sur les thèmes qui s'y rattachent.

Comme nous l'avons vu, une régression de Gauss-Newton typique pour les tests de contraintes peut s'écrire comme

$$\hat{\mathbf{u}} = \hat{\mathbf{X}}\mathbf{b} + \hat{\mathbf{Z}}\mathbf{c} + \text{résidus}, \quad (11.60)$$

où  $\hat{\mathbf{X}}$  est une matrice de dimension  $n \times k$  composée des dérivées de la fonction de régression  $\mathbf{x}(\boldsymbol{\beta})$  évaluées en  $\hat{\boldsymbol{\beta}}$ , les estimations qui satisfont les contraintes et qui sont convergentes au taux  $n^{1/2}$ , et  $\hat{\mathbf{Z}}$  est une matrice de dimension  $n \times r$  composée des régresseurs de test. Dans la plupart des cas que nous examinons,  $\hat{\boldsymbol{\beta}}$  est égal à  $\tilde{\boldsymbol{\beta}}$ , le vecteur des estimations NLS contraintes, auquel cas  $\hat{\mathbf{u}}^\top \hat{\mathbf{X}} = \tilde{\mathbf{u}}^\top \tilde{\mathbf{X}} = \mathbf{0}$ . Cependant, comme il n'y a aucun avantage pour l'intérêt de cette section à faire l'hypothèse la plus forte, nous ne supposons pas que  $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ .

Le numérateur du  $F$  de Fisher pour  $\mathbf{c} = \mathbf{0}$  est égal à la somme des carrés expliqués de la régression

$$\hat{\mathbf{M}}_X \hat{\mathbf{u}} = \hat{\mathbf{M}}_X \hat{\mathbf{Z}}\mathbf{c} + \text{résidus}. \quad (11.61)$$

Si  $\hat{s}^2$  est l'estimation OLS de la variance de (11.60), la statistique de test est  $1/r$  fois

$$\frac{1}{\hat{s}^2} \hat{\mathbf{u}}^\top \hat{\mathbf{M}}_X \hat{\mathbf{Z}} (\hat{\mathbf{Z}}^\top \hat{\mathbf{M}}_X \hat{\mathbf{Z}})^{-1} \hat{\mathbf{Z}}^\top \hat{\mathbf{M}}_X \hat{\mathbf{u}}. \quad (11.62)$$

Le second facteur est ici la somme des carrés expliqués de (11.61). L'expression (11.62) montre clairement que ce que nous testons en réalité, c'est la nullité asymptotique du vecteur à  $r$  composantes,

$$n^{-1/2} \mathbf{Z}^\top \mathbf{M}_X \mathbf{u}. \quad (11.63)$$

Si  $E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}$ , la matrice de covariance asymptotique de ce vecteur est

$$\sigma^2 \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z} \right). \quad (11.64)$$

Puisque (11.62) est une forme quadratique du vecteur (11.63), mais également une quantité qui donne une estimation convergente de sa matrice de covariance, il est aisé de voir qu'elle aura une distribution asymptotique du  $\chi^2(r)$  sous l'hypothèse nulle.

Considérons à présent ce qu'il advient en présence d'hétéroscédasticité. En particulier, supposons que

$$E(\mathbf{u}\mathbf{u}^\top) = \mathbf{\Omega}, \quad (11.65)$$

où  $\mathbf{\Omega}$  est une matrice diagonale dont les éléments diagonaux sont des  $\omega_t^2$  qui satisfont la condition

$$\omega_{\min}^2 < \omega_t^2 < \omega_{\max}^2 \quad \forall t,$$

où  $\omega_{\min}^2$  et  $\omega_{\max}^2$  sont des bornes inférieure et supérieure positives finies. Cette condition élimine la possibilité d'une croissance ou d'une décroissance infinies de  $\omega_t^2$  lorsque  $t \rightarrow \infty$ . Il est évident que si nous n'avons aucune information sur les  $\omega_t^2$ , il nous sera impossible d'en donner une estimation convergente, puisqu'il y aura un  $\omega_t^2$  à estimer pour chaque observation. Néanmoins, il reste possible d'obtenir des estimations convergentes de quantités telles que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{W}^\top \mathbf{\Omega} \mathbf{W} \right), \quad (11.66)$$

où  $\mathbf{W}$  est une matrice composée de  $n$  lignes qui satisfait la condition nécessaire à l'existence de (11.66). Le moyen le plus simple d'obtenir de telles estimations est de faire usage de l'estimateur

$$\frac{1}{n} \mathbf{W}^\top \mathbf{\acute{\Omega}} \mathbf{W},$$

où  $\mathbf{\acute{\Omega}}$  est une matrice diagonale dont l'élément diagonal  $t$  est  $\hat{u}_t^2$ . Ce résultat fondamental est dû à Eicker (1963, 1967) et White (1980). Il permet d'obtenir des matrices de covariance estimées et des statistiques de test qui sont valables malgré une hétéroscédasticité de forme inconnue. Nous démontrerons ce résultat et discuterons des **estimateurs de matrice de covariance robustes à l'hétéroscédasticité**, ou **HCCME**, au cours du Chapitre 16. Pour l'instant,

nous en faisons seulement un usage pour construire des statistiques de test basées sur une régression artificielle.

Si la matrice de covariance de  $\mathbf{u}$  est donnée par (11.65), la matrice de covariance asymptotique du vecteur (11.63) sera

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\Omega} \dot{\mathbf{M}}_X \dot{\mathbf{Z}} \right). \quad (11.67)$$

Grâce au résultat d'Eicker et White, on peut l'estimer de façon convergente par

$$\frac{1}{n} \dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\Omega} \dot{\mathbf{M}}_X \dot{\mathbf{Z}} = \frac{1}{n} \dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{U}} \dot{\mathbf{U}} \dot{\mathbf{M}}_X \dot{\mathbf{Z}},$$

où  $\dot{\mathbf{U}}$  est une matrice diagonale de dimension  $n \times n$  avec  $\dot{u}_t$  comme  $t^{\text{ième}}$  élément diagonal. Par conséquent, la statistique de test

$$\begin{aligned} & \dot{\mathbf{u}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{Z}} (\dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{U}} \dot{\mathbf{U}} \dot{\mathbf{M}}_X \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{u}} \\ &= \boldsymbol{\iota}^\top \dot{\mathbf{U}} \dot{\mathbf{M}}_X \dot{\mathbf{Z}} (\dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{U}} \dot{\mathbf{U}} \dot{\mathbf{M}}_X \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{U}} \boldsymbol{\iota}, \end{aligned} \quad (11.68)$$

où, comme d'habitude,  $\boldsymbol{\iota}$  est un vecteur dont chaque composante égale 1, doit être asymptotiquement distribuée selon le  $\chi^2(r)$  sous l'hypothèse nulle. On peut calculer la valeur de cette statistique comme la somme des carrés expliqués de la régression artificielle

$$\boldsymbol{\iota} = \dot{\mathbf{U}} \dot{\mathbf{M}}_X \dot{\mathbf{Z}} \mathbf{c} + \text{résidus}, \quad (11.69)$$

c'est-à-dire  $n$  moins sa SSR. Nous nous référerons à cette régression en tant que **régression de Gauss-Newton robuste à l'hétéroscédasticité**, ou **HRGNR**, puisque la statistique de test (11.68) est une statistique de test **robuste à l'hétéroscédasticité**.

Bien évidemment, personne ne calcule en réalité la matrice  $\dot{\mathbf{U}}$  dans la pratique, dans le but d'exécuter une HRGNR. Au lieu de cela, on peut procéder comme suit:

- (i) Régresser chaque colonne de  $\dot{\mathbf{Z}}$  sur  $\dot{\mathbf{X}}$  et conserver la matrice des résidus  $\dot{\mathbf{M}}_X \dot{\mathbf{Z}}$ .
- (ii) Multiplier l'élément  $t$  de chaque vecteur de résidus par  $\dot{u}_t$ .
- (iii) Régresser le vecteur  $\boldsymbol{\iota}$  sur les  $r$  régresseurs créés en (ii). Cela correspond à la régression (11.69).
- (iv) Calculer la statistique de test,  $n - \text{SSR}$ . Elle sera asymptotiquement distribuée selon une loi du  $\chi^2(r)$  sous  $H_0$ .

Il se révèle donc être remarquablement simple de calculer un test robuste à l'hétéroscédasticité que l'on peut employer dans les mêmes circonstances que les statistiques de test basées sur la GNR. Pour plus de précisions, consulter Davidson et MacKinnon (1985b), Wooldridge (1990a, 1990b, 1991a), et

MacKinnon (1992). Nous discuterons de la HRGNR plus amplement dans le Chapitre 16.

On devrait mettre l'accent, bien sûr, sur le fait que les résultats théoriques sur lesquels la statistique de test (11.68) repose ne sont vrais qu'asymptotiquement. Bien que cela reste valable également pour les statistiques de test basées sur la GNR, il est presque certainement plus difficile d'estimer la matrice de covariance (11.67) que la matrice de covariance (11.64). Ainsi il faut s'attendre à ce que les tests robustes à l'hétéroscédasticité se comportent moins bien que les tests ordinaires, avec un échantillon fini. Cependant, il y a quelques preuves que les tests basés sur la HRGNR tendent à rejeter l'hypothèse nulle trop peu souvent, en particulier avec le niveau d'erreur de première espèce 0.01; voir Davidson et MacKinnon (1985b).

Dans la pratique, il est sage d'utiliser des tests basés à la fois sur la GNR et sur la HRGNR. Si des tests contre la même hypothèse alternative produisent des résultats similaires, on peut sûrement leur faire confiance. Si ce n'est pas le cas, on désirera sans doute tester, et peut-être transformer le modèle pour en tenir compte, des formes plausibles de l'hétéroscédasticité. On ne devrait jamais avoir confiance en des tests basés sur la GNR si la HRGNR produit des résultats vraiment différents.

## 11.7 CONCLUSION

Au cours de ce chapitre, comme dans les Chapitres 6 et 10, nous avons vu que la régression de Gauss-Newton et sa variante robuste à l'hétéroscédasticité offrent des moyens très simples de tester un grand nombre d'aspects de la spécification d'un modèle pour les modèles de régression. Cependant, nous n'avons rien dit sur la manière d'interpréter les résultats de ces tests et d'autres tests de spécification. C'est le sujet du prochain chapitre.

## TERMES ET CONCEPTS

changement de régime	test de Chow
emboîtement artificiel	test de spécification de la
estimateur de la matrice de covariance	différentiation
robuste à l'hétéroscédasticité	test en $J$
(HCCME)	test en $J_A$ test en $P$
fonction scédastique	test en $P_A$ test robuste à
matrices des différences	l'hétéroscédasticité
modèles emboîtés	tests d'hétéroscédasticité
modèles non emboîtés	tests d'hypothèses non emboîtées
produit direct	tests DWH
régression de Gauss-Newton robuste à	variables instrumentales (IV) (tests de
l'hétéroscédasticité (HRGNR)	modèles estimés par)
sélection de modèle	vecteur de contraste

# Chapitre 12

## Interprétation des Tests Orientés Régression

### 12.1 INTRODUCTION

Dans les chapitres précédents, nous avons discuté d'un grand nombre de statistiques de test pour les modèles de régression linéaire et non linéaire. La plupart de ces tests étaient **orientés régression**, c'est-à-dire qu'il s'agissait de tests de spécification de la fonction de régression. L'usage du terme "orienté" dans ce contexte peut paraître étrange *a priori*, mais il devrait se justifier au fur et à mesure que le chapitre se déroulera. Fondamentalement, les tests orientés régression sont des tests de la spécification de la fonction de régression, alors que les tests **orientés non-régression** sont des tests destinés à d'autres aspects de la modélisation, comme par exemple des tests d'hétéroscédasticité.

Il est désormais temps de connaître la signification des résultats des tests d'hypothèses et la manière de les interpréter. Cette discussion nécessite un certain appareillage technique, et en particulier le concept de **dérive de DGP**, que nous introduirons dans la Section 12.3. L'ensemble des résultats issus de cet appareillage est malgré tout extrêmement simple et intuitif, et il peut être d'une grande utilité dans l'interprétation des statistiques de test que l'on obtient concrètement dans les travaux empiriques. Dans ce chapitre, nous ne discutons que des tests orientés régression pour des modèles de régression estimés par NLS. Bien que cela soit limitatif, cela simplifie considérablement l'exposé. Au cours du prochain chapitre, nous discuterons à la fois des tests de modèles en dehors de la classe des régressions et des tests de modèles de régression dans des directions de non-régression, dans le contexte des trois tests classiques basés sur l'estimation ML, à savoir les tests de Wald, LR et LM. Comme nous le verrons, les principaux résultats de ce chapitre sont transposables sans modification au cas plus général. Ils le sont également, avec quelques remaniements, à des modèles estimés par IV et par GLS.

Dans la Section 3.4, nous introduisons les concepts de niveau et de puissance d'un test. Le niveau d'un test, comme nous le rappellerons, est la probabilité qu'il rejette l'hypothèse nulle lorsque celle-ci est exacte, alors que la puissance d'un test est la probabilité qu'il rejette l'hypothèse nulle lorsque celle-ci est inexacte. A l'évidence, la puissance dépendra de la manière dont les

données auront été générées. Ainsi nous ne pouvons pas parler de puissance sans spécifier un processus générateur de données (ou éventuellement une famille de DGP). En général, la puissance d'un test dépendra de l'hypothèse nulle,  $H_0$ , de l'hypothèse alternative contre laquelle elle est testée,  $H_1$ , et du DGP qui est supposé avoir généré les données. Nous discuterons de certains concepts connexes au niveau et à la puissance des tests dans la Section 12.2.

La puissance d'un test peut dépendre des détails de la construction du test, mais cela ne sera pas important si nous ne nous intéressons qu'aux analyses asymptotiques. De nombreux tests sont **asymptotiquement équivalents** sous l'hypothèse nulle et sous toutes les dérives de DGP, bien qu'ils puissent différer substantiellement avec des échantillons finis. Deux tests sont dits asymptotiquement équivalents s'ils tendent vers la même variable aléatoire. Par exemple, les tests en  $F$  et du  $\chi^2$  basés sur la même régression de Gauss-Newton seront asymptotiquement équivalents, à condition bien sûr que le test en  $F$  soit multiplié par le nombre de degrés de liberté de son numérateur. Ces tests seront également équivalents aux tests en  $F$  ou du  $\chi^2$  asymptotiques contre la même alternative basés sur la comparaison des sommes des résidus au carré des modèles contraint et non contraint.<sup>1</sup> Nous n'essaierons pas de démontrer ce résultat ici; c'est une conséquence de résultats plus généraux démontrés par Davidson et MacKinnon (1987). Cependant, c'est un résultat important, parce qu'il nous permet l'étude des seuls tests basés sur la GNR pour affirmer que nos résultats sont beaucoup plus généralement applicables. Alors, dans ce chapitre, nous discuterons de façon explicite ce qui détermine la puissance asymptotique des tests orientés régression basés sur la GNR, et de façon implicite ce qui détermine la puissance asymptotique de tous les tests orientés régression.

On peut écrire l'hypothèse nulle sous la forme

$$H_0: \mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}. \quad (12.01)$$

Soit  $\tilde{\boldsymbol{\beta}}$  le vecteur à  $k$  paramètres des estimations NLS de  $\boldsymbol{\beta}$ . Alors plusieurs statistiques de test équivalentes peuvent être calculées avec la GNR

$$\mathbf{y} - \tilde{\mathbf{x}} = \tilde{\mathbf{X}}\mathbf{b} + \tilde{\mathbf{Z}}\mathbf{c} + \text{résidus}, \quad (12.02)$$

où, comme d'habitude,  $\tilde{\mathbf{x}}$  désigne  $\mathbf{x}(\tilde{\boldsymbol{\beta}})$ , et où la matrice  $\tilde{\mathbf{X}} \equiv \mathbf{X}(\tilde{\boldsymbol{\beta}})$  de dimension  $n \times k$  a pour élément type  $\partial x_t(\boldsymbol{\beta})/\partial \beta_i$ , et est évaluée en  $\tilde{\boldsymbol{\beta}}$ . Comme nous l'avons vu, la matrice  $\tilde{\mathbf{Z}} \equiv \mathbf{Z}(\tilde{\boldsymbol{\beta}})$  de dimension  $n \times r$  peut être spécifiée de différentes façons, qui dépendent de l'alternative contre laquelle nous voulons

<sup>1</sup> Tous ces tests sont également asymptotiquement équivalents à des tests basés sur la régression de Gauss-Newton robuste à l'hétéroscédasticité discutée dans la Section 11.6, mais uniquement s'il n'y a pas d'hétéroscédasticité. Consulter l'article de Davidson et MacKinnon (1985b).



tester l'hypothèse nulle. La possibilité la plus simple est que  $\mathbf{x}(\boldsymbol{\beta})$  soit un cas particulier de  $\mathbf{x}(\boldsymbol{\beta}, \boldsymbol{\gamma})$  où  $\boldsymbol{\gamma} = \mathbf{0}$ , ce qui nous permet d'écrire

$$H_1: \mathbf{y} = \mathbf{x}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{u}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}. \quad (12.03)$$

Dans ce cas,  $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}_{\boldsymbol{\gamma}}$ , où  $\tilde{\mathbf{X}}_{\boldsymbol{\gamma}}$  a pour élément type  $\partial x_t(\boldsymbol{\beta}, \boldsymbol{\gamma})/\partial \gamma_j$ , évaluée en  $(\tilde{\boldsymbol{\beta}}, \mathbf{0})$ . Cependant, comme nous l'avons vu dans le Chapitre 11, la construction d'un test contre une alternative explicite telle que (12.03) n'est qu'un des nombreux moyens de générer un test basé sur la GNR (12.02).

La statistique de test la plus simple basée sur (12.02) est

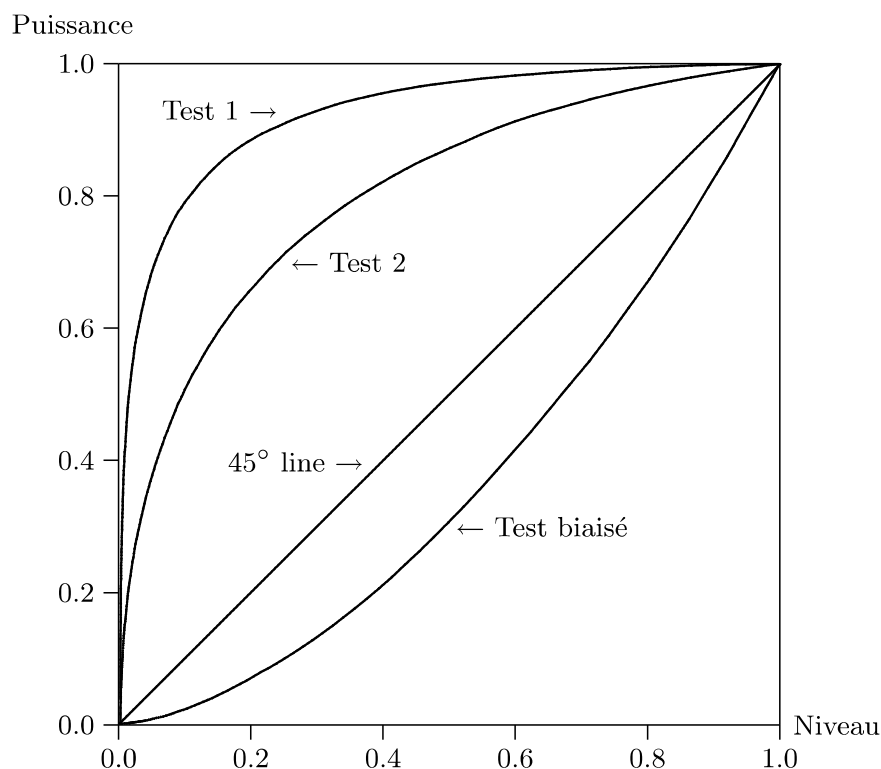
$$\frac{1}{\tilde{s}^2} (\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{Z}} (\tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}^\top (\mathbf{y} - \tilde{\mathbf{x}}), \quad (12.04)$$

où  $\tilde{\mathbf{M}}_X \equiv \mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top$  et  $\tilde{s}^2 \equiv (\mathbf{y} - \tilde{\mathbf{x}})^\top (\mathbf{y} - \tilde{\mathbf{x}})/(n - k)$ . La statistique de test (12.04) est  $1/\tilde{s}^2$  fois la somme des carrés expliqués de (12.02). Par souci de simplicité, nous ne considérerons que cette statistique de test tout au long de ce chapitre. Parce que (12.04) est asymptotiquement équivalente aux autres tests basés sur (12.02) mais aussi aux tests contre la même alternative basés sur les principes de Wald, LR et LM, nos résultats restent malgré tout assez généraux.

Au delà de la spécification de l'hypothèse nulle (12.01) et de la statistique de test (12.04), il nous faut détailler la façon dont nous supposons que les données ont été générées si nous avons l'intention de discuter de la puissance d'un test. Cela nous conduit à considérer le nouveau concept important de dérive de DGP, que nous avons déjà mentionné. Sans ce concept, il serait extrêmement difficile d'analyser les propriétés asymptotiques des statistiques de test lorsque l'hypothèse nulle n'a pas généré les données, et nous discutons donc largement la dérive des DGP dans la Section 12.3. Dans les deux sections qui suivent, nous analysons les propriétés asymptotiques de la statistique de test (12.04) sous certaines dérives de DGP et donnons une interprétation géométrique de ces résultats. Dans la Section 12.6, nous expliquerons comment on pourrait comparer la puissance des tests dont les distributions ne sont connues qu'asymptotiquement. Dans la Section 12.7, nous exploitons les résultats obtenus précédemment et discutons de l'interprétation des résultats des tests orientés régression qui rejettent l'hypothèse nulle. Enfin, dans la Section 12.8, nous verrons comment il faut interpréter les résultats des tests qui ne rejettent pas l'hypothèse nulle.

## 12.2 NIVEAU ET PUISSANCE

Nous avons introduit les concepts de niveau et de puissance des tests d'hypothèses lors de la Section 3.4. Un moyen de voir comment s'articulent ces concepts est d'étudier la **courbe de niveau-puissance** pour n'importe quel test



**Figure 12.1** Courbes de niveau-puissance

donné. Pour simplifier, considérons la statistique de test qui est toujours un nombre positif (les statistiques de test qui sont asymptotiquement distribuées suivant une Fisher ou une  $\chi^2$  possèdent cette propriété). Si nous choisissons une valeur critique nulle, le test rejettera constamment l'hypothèse nulle, que le DGP soit véritablement un cas particulier de l'hypothèse nulle ou pas. Au fur et à mesure que nous augmentons la valeur critique, la probabilité que le test rejette l'hypothèse nulle décroît. Si le test est utile, cette probabilité diminuera à l'origine beaucoup moins rapidement lorsque l'hypothèse nulle est fausse que lorsqu'elle est vraie. La courbe de niveau-puissance montre, pour une taille d'échantillon donnée, ces deux probabilités simultanément. L'axe des abscisses est celui du niveau calculé pour un DGP qui satisfait l'hypothèse nulle, et l'axe des ordonnées est celui de la puissance, pour un autre DGP donné qui ne satisfera pas en général l'hypothèse nulle. Ainsi la courbe de niveau-puissance illustre ce qu'est la puissance du test contre le DGP donné pour chaque niveau de test que l'on peut choisir.

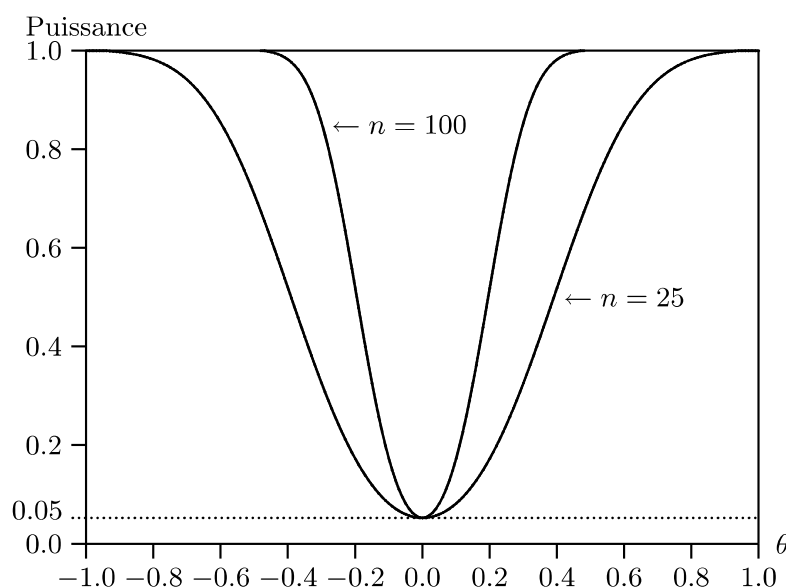
Considérons à présent la Figure 12.1, qui illustre quelques courbes de niveau-puissance pour différentes statistiques de test potentielles. L'axe des abscisses mesure le niveau. L'axe des ordonnées mesure la puissance, lorsque les données sont générées par un DGP fixé. La courbe de niveau-puissance est générée en modifiant la valeur critique du test. L'angle du nord-est correspond à une valeur critique nulle. La puissance et le niveau sont, en ce point,

unitaires. L'angle du sud-ouest correspond à une valeur critique très grande, et tellement élevée que la statistique de test ne lui sera jamais supérieure. La puissance et le niveau sont, en ce point, égaux à 0. Pour de nombreuses statistiques de test, dont celles distribuées selon une  $\chi^2$  sous l'hypothèse nulle, cette valeur critique est en principe infinie. Cependant, nous pourrions sélectionner une valeur critique finie telle que la statistique de test la dépasse avec une probabilité aussi proche de 0 que l'on veut.

La courbe de niveau-puissance d'un test pour lequel le niveau égale la puissance correspond à la première bissectrice. Cela sera le cas par définition si le DGP pour lequel la courbe est construite satisfait véritablement l'hypothèse nulle. En dehors de ce cas, un test qui donnerait ce résultat serait à l'évidence peu utile. Normalement, nous nous attendons à ce que la puissance d'un test soit supérieure à son niveau pour n'importe quelle valeur critique, excepté dans le cas où le niveau et la puissance sont égaux soit à 1 soit à 0. Les courbes désignées "Test 1" et "Test 2" sur la figure sont des exemples de tests pour lesquels c'est le cas. Cependant, il existe des tests pour lesquels le niveau est supérieur à la puissance pour certains DGP. Ces tests sont appelés **tests biaisés**, et la courbe appelée "Test Biaisé" illustre ce phénomène. Pour une discussion plus profonde sur les tests biaisés, qui sont évidemment très peu utilisés, consulter Kendall et Stuart (1979, Chapitre 23).

Il est clair à partir de la Figure 12.1 que le Test 1 est plus utile que le Test 2. À l'exception des deux extrémités, la courbe de niveau-puissance pour le premier est partout au dessus de la courbe du second. Ainsi, pour n'importe quel niveau, la puissance du Test 1 est plus forte que celle du Test 2. Si la taille augmente, nous nous attendons à ce que la courbe de niveau-puissance d'un test qui a de bonnes propriétés s'améliore (c'est-à-dire qu'elle s'éloigne de la première bissectrice). À la limite, lorsque  $n \rightarrow \infty$ , la courbe de niveau-puissance ressemblerait à  $\Gamma$ , passant par les points  $(0, 0)$ ,  $(0, 1)$ , et  $(1, 1)$ .

On peut générer des courbes de niveau-puissance avec ce que l'on appelle la **fonction puissance** d'un test. Cette fonction fournit la puissance d'un test comme fonction de son niveau (ou de façon équivalente, de la valeur critique), de la taille de l'échantillon, et du DGP. Habituellement, le DGP est contraint à appartenir à une hypothèse alternative particulière caractérisée par un ensemble fini de paramètres. Spanos (1986, Chapitre 14) donne une définition formelle des fonctions puissance dans ce contexte. Supposons, pour être concrets, que nous nous intéressions à un unique paramètre  $\theta$  et que l'hypothèse nulle soit  $\theta = 0$ . Lorsque  $\theta = 0$ , la puissance du test sera bien évidemment égale à son niveau. Pour toute autre valeur de  $\theta$ , la puissance sera supérieure au niveau si le test est sans biais. Pour un test possédant de bonnes propriétés, nous espérons que, pour une taille d'échantillon raisonnable, la puissance augmentera de façon monotone avec  $|\theta|$  et convergera vers 1 lorsque  $|\theta| \rightarrow \infty$ . De façon similaire, pour tout  $\theta \neq 0$ , nous nous attendons à ce que la puissance tende vers 1 lorsque la taille de l'échantillon tend vers l'infini. La Fi-



**Figure 12.2** Fonctions puissance pour tests de  $\theta = 0$  au niveau de .05

Figure 12.2 illustre deux fonctions puissance, pour un test identique mais des tailles d'échantillon différentes. Les données sont générées à partir de la loi  $N(\theta, 1)$ , et l'hypothèse nulle est  $\theta = 0$ . Les fonctions puissance sont illustrées pour des tests à un taux de 5% avec des tailles d'échantillon égales à 25 et 100. Ces fonctions puissance sont symétriques par rapport à 0. Comme nous l'espérons, la fonction puissance pour  $n = 100$  est partout supérieure à la fonction puissance pour  $n = 25$ , sauf en  $\theta = 0$ .

Si un test rejette une hypothèse nulle fausse avec une probabilité asymptotiquement nulle, on parle de test **convergent**. Le concept de convergence pour un test fut introduit par Wald et Wolfowitz (1940). C'est un concept simple et intuitif et c'est évidemment une propriété recherchée pour un test. Le test illustré sur la Figure 12.2 est convergent. Par conséquent, lorsque  $n \rightarrow \infty$ , la fonction puissance tend vers la forme d'un  $\top$ , avec une puissance égale à 1 pour toute valeur de  $\theta$  sauf  $\theta = 0$ . Nous pouvons définir la **convergence** d'un test d'hypothèses de façon formelle comme suit.

*Définition 12.1.*

Un test est convergent contre une certaine classe de DGP dont aucun ne satisfait l'hypothèse nulle si, lorsque les données sont générées par un membre appartenant à cette classe, la probabilité de rejeter l'hypothèse nulle tend vers 1 lorsque la taille de l'échantillon  $n$  tend vers l'infini, pour n'importe quelle valeur critique associée à un niveau non nul.

Remarquons que la propriété de convergence d'un test dépendra de la façon dont sont générées les données. Un test qui est convergent contre certains DGP peut ne pas l'être contre d'autres. Intuitivement, la raison pour la-

quelle les tests sont souvent convergents est que lorsque  $n \rightarrow \infty$ , la masse d'informations portée par l'échantillon sur la validité de l'hypothèse nulle s'accroît sans limite. Ce faisant, l'information étouffe le bruit des données et permet finalement de conclure avec une probabilité égale à 1 que la statistique de test *n'est pas* un tirage de ce qui serait sa distribution sous l'hypothèse nulle.

Ces préliminaires étant faits, nous pouvons considérer ce qui détermine la puissance des tests orientés régression. Puisque nous traitons des modèles de régression non linéaire, il nous faut nous baser sur une analyse asymptotique. Cependant, cela soulève une difficulté technique de taille. Tous les tests considérés jusqu'à présent sont convergents lorsque les données sont générées par un DGP fixé appartenant à l'ensemble des alternatives, et ils sont en réalité plus convergents que cela. Si un test est convergent, la valeur de la statistique de test tendra vers plus ou moins l'infini lorsque  $n \rightarrow \infty$ . Cela nous empêche de parler de la **distribution asymptotique** d'une telle statistique de test, mais aussi de comparer les distributions asymptotiques de deux statistiques concurrentes lorsque les deux tests sont convergents, si le DGP est fixé. La solution consiste à laisser dériver un DGP vers l'hypothèse nulle à un certain taux. C'est dans la prochaine section que nous parlons de dérive de DGP.

## 12.3 DÉRIVE DE DGP

Afin de déterminer les propriétés d'une statistique de test, il faut spécifier le processus qui génère les données. Puisque, dans ce chapitre, nous ne nous intéressons qu'aux tests orientés régression, nous focaliserons nos efforts sur les DGP qui ne diffèrent de l'hypothèse nulle que dans ces directions. Cette limitation n'est en aucune manière anodine. Elle signifie que nous ne pouvons rien dire sur la puissance des tests orientés régression lorsque le modèle est mal spécifié ailleurs que dans la fonction de régression (par exemple, lorsque les aléas sont sujets à une hétéroscédasticité non modélisée). Certains aspects de ce thème seront abordés lors du Chapitre 16.

La manière naturelle de spécifier un DGP dans le but d'analyser la puissance d'un test consiste à supposer que c'est un cas particulier de la classe des DGP qui composent ensemble l'hypothèse alternative. Cependant, on note deux problèmes relatifs à cette approche. En premier lieu, on peut parfaitement s'intéresser à la puissance de certains tests lorsque les données sont générées par un DGP qui n'appartient pas à l'hypothèse alternative. Il semble peu pertinent d'éliminer d'office ce cas intéressant.

Le second problème, auquel nous avons fait allusion dans la section précédente est que la plupart des statistiques de test qui nous intéressent ne posséderont pas de distribution asymptotique non dégénérée sous un DGP fixé qui n'est pas un cas particulier de l'hypothèse nulle. Si c'était le cas, elles ne seraient pas convergentes. Une solution éprouvée serait de considérer la

distribution de la statistique de test à laquelle nous nous intéressons sous ce que l'on nomme une **suite d'alternatives locales**. Lorsque  $\boldsymbol{\theta}$  est le vecteur de paramètres d'intérêt, on peut écrire une suite de ce type comme

$$\boldsymbol{\theta}^n = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta}. \quad (12.05)$$

Ici  $\boldsymbol{\theta}^n$  est le vecteur de paramètres correspondant à une taille d'échantillon égale à  $n$ ,  $\boldsymbol{\theta}_0$  est un vecteur de paramètres qui satisfait l'hypothèse nulle, et  $\boldsymbol{\delta}$  est un vecteur non nul. A l'évidence,  $\boldsymbol{\theta}^n$  converge vers  $\boldsymbol{\theta}_0$  à un taux proportionnel à  $n^{-1/2}$ . Le pionnier de cette approche est Neyman (1937). Cependant, on l'attribue souvent à Pitman (1949) et on s'y réfère souvent sous le nom de “suite de Pitman” ou “dérive de Pitman”; voir McManus (1991). Cette technique a été abondamment employée en économétrie; voir, par exemple, Gallant et Holly (1980) et Engle (1984).

Afin de ne pas éliminer le cas intéressant où les données sont générées par un DGP qui n'appartient pas à l'hypothèse alternative, Davidson et MacKinnon (1985a, 1987) ont généralisé l'idée de suites d'alternatives locales à l'idée de dérive de DGP. Ce chapitre s'inspire largement de l'approche initiée par les deux articles.<sup>2</sup>

Une classe de dérive de DGP adéquate pour l'étude de la puissance de la statistique de test (12.04) est

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0) + \alpha n^{-1/2}\mathbf{a} + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma_0^2\mathbf{I}. \quad (12.06)$$

Ici  $\boldsymbol{\beta}_0$  et  $\sigma_0^2$  désignent des valeurs spécifiques pour  $\boldsymbol{\beta}$  et  $\sigma^2$ ,  $\mathbf{a}$  est un vecteur à  $n$  composantes qui peut dépendre de variables exogènes, du vecteur de paramètres  $\boldsymbol{\beta}_0$ , et éventuellement des valeurs passées de  $y_t$ , et  $\alpha$  est un paramètre qui détermine la distance séparant le DGP de **hypothèse nulle simple**

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0) + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma_0^2\mathbf{I}. \quad (12.07)$$

La dérive de DGP (12.06) tend vers l'hypothèse nulle lorsque  $n \rightarrow \infty$ . Nous discuterons bientôt du sens précis à donner au vecteur  $\mathbf{a}$ . Remarquons que, lorsque  $n$  croît, le vecteur  $\mathbf{y}$  se rapproche de ce qu'il devrait être sous l'hypothèse nulle simple (12.07) à un taux proportionnel à  $n^{-1/2}$ .

Le fait que la dérive de DGP (12.06) converge vers l'hypothèse nulle simple (12.07) à un taux de  $n^{-1/2}$  n'est pas un hasard. Ce taux a été choisi avec minutie de manière à ce que la statistique de test (12.04), et toutes les statistiques de test asymptotiquement équivalentes, aient une distribution asymptotique lorsque  $n \rightarrow \infty$ . De façon similaire, pour un niveau de test fixé, la valeur de la fonction puissance tend vers une limite qui n'est en général ni

<sup>2</sup> Pour être exact, le terme employé par Davidson et MacKinnon (1985a, 1987) était “suite de DGP locaux”. Cependant, notre préférence va désormais au terme “dérive de DGP”.

0 ni 1 lorsque  $n \rightarrow \infty$  et lorsque la dérive de DGP converge vers l'hypothèse nulle au taux  $n^{-1/2}$ . Cette fonction limite s'appelle **fonction puissance asymptotique** de la statistique de test.

La dérive de DGP (12.06) fournit une représentation *locale* parfaitement générale de tout modèle de régression suffisamment proche de (12.07). Supposons, par exemple, que l'on veuille connaître le comportement d'un test lorsque les données sont générées par une alternative telle que (12.03), où  $\gamma \neq \mathbf{0}$ . Nous pourrions spécifier une suite d'alternatives locales comme

$$\mathbf{y} = \mathbf{x}(\beta_0, \alpha n^{-1/2} \gamma_0) + \mathbf{u}, \quad (12.08)$$

où  $\gamma_0$  est fixé et peut être normalisé à une longueur arbitraire, et où  $\alpha$  détermine la distance qui sépare (12.08) de (12.07). Parce que (12.08) converge vers (12.07) au même taux que  $n^{-1/2}$  converge vers 0, un développement en série de Taylor au premier ordre de (12.08) autour de  $\alpha = 0$  doit donner exactement les mêmes résultats, dans une analyse asymptotique, que (12.08) elle-même. Cette approximation est

$$\mathbf{y} = \mathbf{x}(\beta_0, \mathbf{0}) + \alpha n^{-1/2} \mathbf{X}_\gamma(\beta_0, \mathbf{0}) \gamma_0 + \mathbf{u}, \quad (12.09)$$

où  $\mathbf{X}_\gamma(\beta_0, \mathbf{0})$  a pour élément type  $\partial x_t(\beta, \gamma) / \partial \gamma_j$  évaluée en  $[\beta_0 \ ; \ \mathbf{0}]$ . Si nous définissons  $\mathbf{x}(\beta_0)$  par  $\mathbf{x}(\beta_0, \mathbf{0})$  et  $\mathbf{a}$  par  $\mathbf{X}_\gamma(\beta_0, \mathbf{0}) \gamma_0$ , nous voyons immédiatement que (12.09) est simplement un cas particulier de la dérive de DGP (12.06).

L'argument précédent devrait montrer clairement que (12.06) est une manière tout à fait générale de spécifier une dérive de DGP correspondant à *n'importe quel* modèle de régression alternatif qui comprend l'hypothèse nulle (12.01). Toute alternative spécifique produit simplement un vecteur  $\mathbf{a}$  différent. Si  $\mathbf{a}$  est un vecteur nul, le DGP est un cas particulier de l'hypothèse nulle, et le test aura une puissance égale à son niveau et par conséquent, aura une courbe de niveau-puissance confondue avec la première bissectrice (voir la Figure 12.1). Si  $\mathbf{a}$  est construit à partir de l'hypothèse alternative contre laquelle le test est fondé, alors la dérive de DGP (12.06) est véritablement une suite d'alternatives locales telle que (12.05). En général, cependant, aucun de ces cas particuliers ne se produira.

## 12.4 DISTRIBUTION ASYMPTOTIQUE DES STATISTIQUES

Nous sommes à présent parés pour trouver la distribution asymptotique de la statistique de test (12.04) sous la famille de dérive de DGP (12.06). Afin de valider notre analyse asymptotique, il nous faut supposer que des conditions de régularité variées sont vérifiées. Ainsi, nous supposons que  $n^{-1} \mathbf{X}_0^\top \mathbf{X}_0$ ,  $n^{-1} \mathbf{Z}_0^\top \mathbf{Z}_0$ , et  $n^{-1} \mathbf{Z}_0^\top \mathbf{X}_0$  sont des matrices qui tendent toutes vers des matrices limites finies de rangs  $k$ ,  $r$  et  $\min(k, r)$  respectivement lorsque  $n \rightarrow \infty$ . Nous

supposerons ensuite qu'il existe un  $N$  tel que, pour tout  $n > N$ , le rang de la matrice  $[\mathbf{X}_0 \ \mathbf{Z}_0]$  est  $k + r$ , que  $n^{-1}\mathbf{a}^\top \mathbf{a}$  tend vers un scalaire fini, et que  $n^{-1}\mathbf{a}^\top \mathbf{X}_0$  et  $n^{-1}\mathbf{a}^\top \mathbf{Z}_0$  tendent vers des vecteurs limites finis de dimensions  $1 \times k$  et  $1 \times r$  respectivement. Ici  $\mathbf{X}_0$  désigne  $\mathbf{X}(\beta_0)$  et  $\mathbf{Z}_0$  désigne  $\mathbf{Z}(\beta_0)$ . La validité des conditions de régularité dépendra du vecteur  $\mathbf{a}$ , de l'hypothèse nulle (12.01), de l'hypothèse alternative (qu'elle soit exacte ou non), et de l'hypothèse nulle simple (12.07).

Nous commençons par écrire la statistique de test (12.04) de façon à ce qu'elle corresponde au produit de quatre facteurs, qui sont tous  $O(1)$ :

$$\frac{1}{\tilde{s}^2} (n^{-1/2}(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{Z}}) (n^{-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}})^{-1} (n^{-1/2} \tilde{\mathbf{Z}}^\top (\mathbf{y} - \tilde{\mathbf{x}})). \quad (12.10)$$

Il nous faut maintenant remplacer les quantités  $\tilde{s}$ ,  $n^{-1/2}(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{Z}}$ , et  $n^{-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}}$  par leur limite asymptotique sous l'hypothèse (12.06). Nous établissons les résultats suivants sans démonstration. Ils s'obtiennent tous par une modification pertinente des arguments invoqués dans le Chapitre 5:

$$\tilde{s}^2 \xrightarrow{p} \sigma_0^2, \quad (12.11)$$

$$n^{-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}} \xrightarrow[n \rightarrow \infty]{p} \text{plim} (n^{-1} \mathbf{Z}_0^\top \mathbf{M}_X \mathbf{Z}_0), \quad (12.12)$$

et

$$n^{-1/2}(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{Z}} \stackrel{a}{=} n^{-1/2}(\mathbf{u} + \alpha n^{-1/2} \mathbf{a})^\top \mathbf{M}_X \mathbf{Z}_0, \quad (12.13)$$

où  $\mathbf{M}_X \equiv \mathbf{I} - \mathbf{X}_0(\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top$ .

L'intuition qui guide les résultats (12.11) et (12.12) est immédiate. La dérive de DGP (12.06) converge vers l'hypothèse nulle simple (12.07) suffisamment vite pour que les limites de  $\tilde{s}^2$  et  $n^{-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}}$  soient exactement les mêmes que sous l'hypothèse (12.07). Ces limites,  $\sigma_0^2$  et  $\text{plim}(n^{-1} \mathbf{Z}_0^\top \mathbf{M}_X \mathbf{Z}_0)$ , sont déterministes parce que la différence entre  $\tilde{\beta}$  et  $\beta_0$ , qui est  $O(n^{-1/2})$ , n'affecte ni  $\tilde{s}^2$  ni  $n^{-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}}$  asymptotiquement. Il est par conséquent peu surprenant que la différence entre la dérive de DGP (12.06) et l'hypothèse nulle simple (12.07), qui est également  $O(n^{-1/2})$ , n'ait aucun effet sur  $\tilde{s}^2$  et sur  $n^{-1} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \tilde{\mathbf{Z}}$  asymptotiquement.

Par contraste,  $n^{-1/2}(\mathbf{y} - \tilde{\mathbf{x}})^\top \tilde{\mathbf{Z}}$  tend vers une limite en probabilité aléatoire. Le résultat (12.13) provient du fait que

$$\mathbf{y} - \tilde{\mathbf{x}} = \mathbf{M}_X(\mathbf{u} + \alpha n^{-1/2} \mathbf{a}) + o(n^{-1/2}),$$

qui est l'analogue du résultat qui nous est familier (5.57) dans le cas où  $\alpha = 0$ . La raison pour laquelle  $\alpha n^{-1/2} \mathbf{a}$  a un impact est que  $\tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \mathbf{u}$  et  $\alpha n^{-1/2} \tilde{\mathbf{Z}}^\top \tilde{\mathbf{M}}_X \mathbf{a}$  sont du même ordre,  $O(n^{1/2})$ . Ainsi, en spécifiant la dérive de DGP (12.06) comme nous l'avons fait, nous garantissons que les quantités qui sont asymptotiquement déterministes sous l'hypothèse nulle simple



(12.07) ne sont pas modifiées sous (12.06), alors que des quantités qui sont asymptotiquement aléatoires le sont.

La substitution de (12.11), (12.12) et (12.13) dans (12.20) nous permet de voir que la statistique de test (12.04) est asymptotiquement égale à

$$\frac{1}{n\sigma_0^2}(\alpha n^{-1/2}\mathbf{a} + \mathbf{u})^\top \mathbf{M}_X \mathbf{Z} \operatorname{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z} \right)^{-1} \mathbf{Z}^\top \mathbf{M}_X (\alpha n^{-1/2}\mathbf{a} + \mathbf{u}), \quad (12.14)$$

où, pour simplifier la notation,  $\mathbf{Z}$  désigne  $\mathbf{Z}_0$ . Il reste à déterminer la distribution asymptotique de cette quantité. Premièrement, nous définissons  $\boldsymbol{\psi}$  comme une matrice triangulaire de dimension  $r \times r$  telle que

$$\boldsymbol{\psi} \boldsymbol{\psi}^\top \equiv \operatorname{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z} \right)^{-1}. \quad (12.15)$$

Nous définissons ensuite  $\boldsymbol{\eta}$  le vecteur de dimension  $r$  tel que

$$\boldsymbol{\eta} \equiv \frac{1}{\sigma_0} \boldsymbol{\psi}^\top \mathbf{Z}^\top \mathbf{M}_X (\alpha n^{-1}\mathbf{a} + n^{-1/2}\mathbf{u}).$$

La quantité (12.14) prend désormais la forme simple  $\boldsymbol{\eta}^\top \boldsymbol{\eta}$ ; il s'agit simplement de la somme de  $r$  variables aléatoires au carré, les  $r$  éléments du vecteur  $\boldsymbol{\eta}$ .

Il est aisé de voir que, asymptotiquement, l'espérance de  $\boldsymbol{\eta}$  est le vecteur

$$\operatorname{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \alpha \sigma_0^{-1} \boldsymbol{\psi}^\top \mathbf{Z}^\top \mathbf{M}_X \mathbf{a} \right) \quad (12.16)$$

et que sa matrice de covariance est

$$\operatorname{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sigma_0^{-2} \boldsymbol{\psi}^\top \mathbf{Z}^\top \mathbf{M}_X E(\mathbf{u} \mathbf{u}^\top) \mathbf{M}_X \mathbf{Z} \boldsymbol{\psi} \right) = \boldsymbol{\psi}^\top \operatorname{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z} \right) \boldsymbol{\psi} = \mathbf{I}_r.$$

La dernière égalité provient ici de la définition de  $\boldsymbol{\psi}$  dans (12.15). Puisque  $\boldsymbol{\eta}$  est égal à la somme d'un terme qui tend vers la limite déterministe (12.16) et de  $n^{-1/2}$  fois une somme pondérée de variables aléatoires de variances finies, et puisque notre hypothèse conserve ces poids à l'intérieur de bornes inférieure et supérieure, nous pouvons appliquer un théorème de la limite centrale. La statistique de test (12.04) est ainsi asymptotiquement égale à une somme de  $r$  variables aléatoires indépendantes normales au carré, toutes de variance unitaire et d'espérance donnée par un élément du vecteur (12.16). Une telle somme suit la **distribution du chi-carré non centrée** à  $r$  degrés de liberté et dont le **paramètre de non centralité**, ou **NCP**, est égal à la norme au carré du vecteur d'espérances (12.16).

La distribution du  $\chi^2$  non centrée joue un rôle majeur dans l'analyse de la puissance asymptotique de la plupart des tests économétriques. Cette distribution est abordée brièvement dans l'Annexe B; pour une discussion plus complète, les lecteurs devraient consulter Johnson et Kotz (1970b,

Chapitre 28). L'allure de cette distribution dépend de deux éléments: le nombre de degrés de liberté et le NCP. Le NCP est toujours un nombre positif; s'il est nul, nous aurions une distribution du  $\chi^2$  centrale ordinaire.

Afin de développer notre intuition, il est révélateur de considérer le cas à deux degrés de liberté. Supposons que  $\varepsilon_1$  et  $\varepsilon_2$  soient des variables aléatoires indépendantes, distribuées selon une  $N(0, 1)$ , et supposons par ailleurs que  $\xi_1 = \mu_1 + \varepsilon_1$  et  $\xi_2 = \mu_2 + \varepsilon_2$ , où  $\mu_1$  et  $\mu_2$  sont des valeurs fixées. La statistique

$$\zeta^C \equiv \varepsilon_1^2 + \varepsilon_2^2$$

sera distribuée suivant une  $\chi^2(2)$ , alors que la statistique

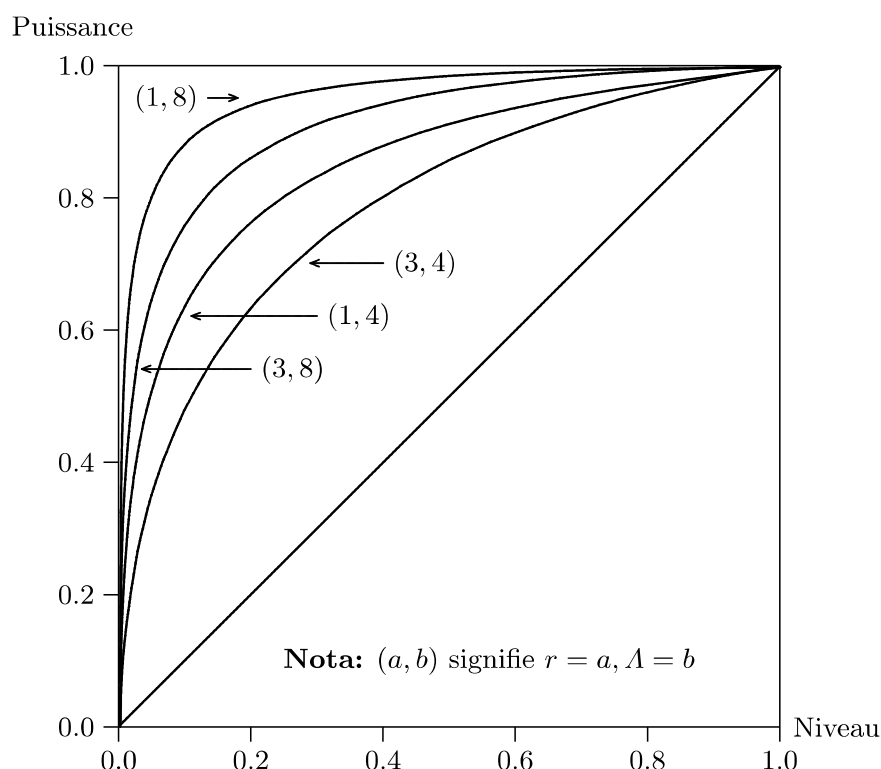
$$\zeta^N \equiv \xi_1^2 + \xi_2^2 = (\varepsilon_1^2 + \varepsilon_2^2) + (\mu_1^2 + \mu_2^2) + (2\mu_1\varepsilon_1 + 2\mu_2\varepsilon_2) \quad (12.17)$$

sera distribuée selon une  $\chi^2(2)$  non centrée et dont le NCP est égal à  $\mu_1^2 + \mu_2^2$ . Une notation standard pour la distribution du  $\chi^2$  non centrée est  $\chi^2(r, \Lambda)$ , où  $r$  est le nombre de degrés de liberté et  $\Lambda$  est le NCP. Ainsi, dans ce cas, nous pourrions dire que  $\zeta^N$  est distribué selon une  $\chi^2(2, \mu_1^2 + \mu_2^2)$ .<sup>3</sup>

L'espérance de  $\zeta^N$  est supérieure à celle de  $\zeta^C$ . Cette dernière est égale à 2, alors que la première est égale à  $2 + \mu_1^2 + \mu_2^2$ . Ainsi, en moyenne,  $\zeta^N$  sera supérieure à  $\zeta^C$ . Donc, si nous devions tester l'hypothèse (erronée) que  $\zeta^N$  provient de la distribution du  $\chi^2(2)$  *centrée* à l'aide d'un test de niveau  $\delta$ , nous rejeterions cette hypothèse dans plus de  $100\delta\%$  des cas. La puissance de ce test, puisque nous conservons un nombre de degrés de liberté constant, ne dépendra que du NCP,  $\mu_1^2 + \mu_2^2$ . Connaissant (12.17), cela peut paraître étrange. Il semblerait que la distribution de  $\zeta^N$  dépende de  $\mu_1$  et de  $\mu_2$  individuellement plutôt que de la somme de leurs carrés. En réalité, les variations de  $\mu_1$  et  $\mu_2$  qui ne modifient pas  $\mu_1^2 + \mu_2^2$  sont sans effet sur la distribution de  $\zeta^N$ . La démonstration serait un bon exercice.

On associe au  $\chi^2$  non centré deux autres distributions, appelées  $F$  non centrée et  $F$  doublement non centrée. Elles sont définies de façon analogue à la distribution en  $F$  ordinaire (centrée), comme un rapport de deux variables aléatoires indépendantes du  $\chi^2$ , divisée chacune par son degré de liberté. Pour la distribution en  $F$  non centrée, la variable aléatoire du numérateur obéit à une distribution du  $\chi^2$  non centrée, alors que celle du dénominateur obéit à une  $\chi^2$  centrée. Pour la distribution en  $F$  doublement non centrée, à la fois le

<sup>3</sup> Remarquons que certains auteurs, et aussi certains logiciels informatiques, utilisent la racine carrée de  $\Lambda$ , plutôt que  $\Lambda$  lui-même, en tant que NCP et se réfèrent donc à cette racine carrée en tant que NCP. La paramétrisation de la non centralité de la distribution du  $\chi^2$  n'a pas d'importance. Cependant, la paramétrisation employée ici est plus naturelle mais aussi plus répandue: si  $x_1 \sim \chi^2(r_1, \Lambda_1)$  et  $x_2 \sim \chi^2(r_2, \Lambda_2)$  sont indépendantes, alors  $z = x_1 + x_2$  est distribuée selon une  $\chi^2(r_1 + r_2, \Lambda_1 + \Lambda_2)$ . Cela devrait illustrer le fait que  $\Lambda$ , plutôt que sa racine carrée, est un choix naturel pour le NCP.



**Figure 12.3** Les courbes de niveau-puissance dépendent de  $r$  et  $\Lambda$

numérateur et le dénominateur ont des distributions du  $\chi^2$  non centrées. Si l'on étudie la puissance d'un test en  $F$  ordinaire dans le modèle de régression à aléas normaux, avec un DGP fixé plutôt qu'une dérive de DGP, on trouve que la statistique de test est distribuée suivant une distribution de Fisher soit non centrée (si le DGP est un cas particulier de l'alternative) soit doublement non centrée (dans le cas contraire). La difficulté supplémentaire de la distribution de Fisher doublement non centrée survient dans le second cas parce qu'il n'implique pas de dérive de DGP. Par conséquent, l'estimation de  $\sigma^2$  sous l'alternative n'est pas d'espérance égale à  $\sigma_0^2$ , ce qui nous empêche de calculer la limite lorsque  $n \rightarrow \infty$ . Alors, à plusieurs titres, l'analyse asymptotique de modèles non linéaires est plus simple que l'analyse de modèles linéaires avec des échantillons finis. Pour une discussion des modèles linéaires, voir Thursby et Schmidt (1977).

Si une statistique de test obéit à une distribution du  $\chi^2(r)$  sous l'hypothèse nulle et obéit à une distribution du  $\chi^2(r, \Lambda)$  sous une dérive de DGP, la puissance du test dépendra uniquement de  $r$  et  $\Lambda$ . En réalité, elle sera strictement croissante en  $\Lambda$  et strictement décroissante en  $r$ ; voir Das Gupta et Perlman (1974). L'espérance de la statistique sera égale à  $r + \Lambda$ . Ainsi, si  $\Lambda$  augmente, la chance de voir la statistique de test dépasser n'importe quelle valeur critique utilisée doit augmenter. A la limite, lorsque  $\Lambda \rightarrow \infty$ , la puissance du test tend vers 1 pour n'importe quelle valeur critique sélectionnée. La

Figure 12.3 illustre la dépendance de la puissance à  $r$  et  $\Lambda$ , et nous observons quatre cas différents de courbes de niveau-puissance. Ces quatre cas, ordonnés par puissances décroissantes pour un niveau donné sont  $(1, 8)$ ,  $(3, 8)$ ,  $(1, 4)$ , et  $(3, 4)$ , où le premier élément de chaque couple est  $r$  alors que le second est  $\Lambda$ .

Revenons à présent à la statistique de test (12.04). Nous avons vu qu'elle est asymptotiquement distribuée selon la  $\chi^2(r, \Lambda)$  avec un paramètre de non centralité  $\Lambda$  égal à la norme au carré de (12.16). Typiquement,

$$\Lambda = \frac{\alpha^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{a}^\top \mathbf{M}_X \mathbf{Z} \right) \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z} \right)^{-1} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_X \mathbf{a} \right). \quad (12.18)$$

Pour un test particulier d'un modèle donné,  $\mathbf{M}_X$ ,  $\mathbf{Z}$ , et  $r$  sont fixés. Le seul élément variable est la dérive de DGP qui est supposée avoir généré les observations. L'étude de (12.18) montre comment le scalaire  $\alpha$  et le vecteur  $\mathbf{a}$  modifient  $\Lambda$  et donc indirectement la puissance du test. Nous voyons immédiatement que  $\Lambda$  est proportionnel à  $\alpha^2$ . Ainsi  $\alpha$  est simplement un paramètre qui mesure la distance entre la dérive de DGP (12.06) et l'hypothèse nulle simple (12.07). A contrario,  $\mathbf{a}$  mesure la *direction* dans laquelle le DGP s'éloigne de l'hypothèse nulle simple (12.07).

Afin de saisir l'essence de (12.18) et ses conséquences pour la puissance d'un test, il est extrêmement révélateur de considérer l'aspect géométrique des choses. C'est ce que nous faisons dans la section qui suit.

## 12.5 LA GÉOMÉTRIE DE LA PUISSANCE DES TESTS

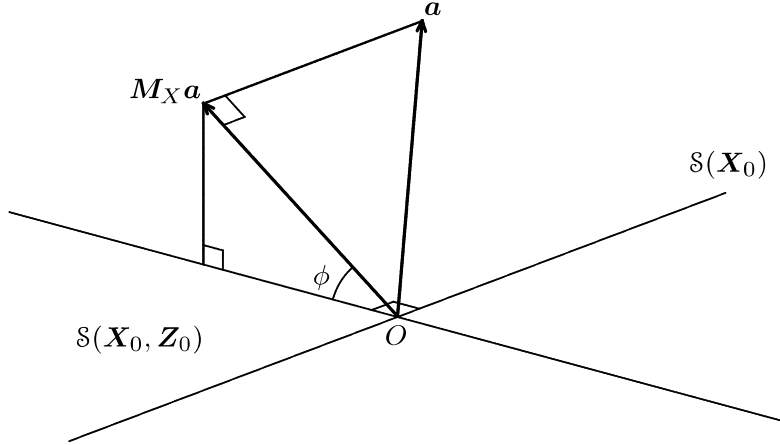
La NCP (12.18) n'est guère parlant sous cette forme. Il est possible, toutefois, de le récrire de façon plus claire. En premier lieu, considérons le vecteur  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{a}$ , dont la longueur au carré est asymptotiquement

$$\alpha^2 \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{a}^\top \mathbf{M}_X \mathbf{a} \right). \quad (12.19)$$

Cette quantité est  $\alpha^2$  fois la limite en probabilité des résidus au carré de la régression de  $n^{-1/2} \mathbf{a}$  sur  $\mathbf{X}_0$ . Supposons que pour une valeur fixée de  $n$  le DGP correspondant à cette taille d'échantillon soit représenté par le vecteur  $\mathbf{x}(\beta_0) + \alpha n^{-1/2} \mathbf{a}$  dans  $E^n$ . Si l'hypothèse nulle est représentée comme dans la Section 2.2 par la variété  $\mathcal{X}$  générée par les vecteurs  $\mathbf{x}(\beta)$  en faisant varier  $\beta$ , la somme des résidus au carré considérée plus haut est le carré de la distance euclidienne entre le point représentant de DGP et l'approximation linéaire  $\mathcal{S}(\mathbf{X}_0)$  à la variété  $\mathcal{X}$  au point  $\beta_0$ . Elle fournit par conséquent une mesure de la différence, pour un  $n$  donné, entre le modèle testé et le DGP.

Considérons à présent la régression artificielle

$$(\alpha/\sigma_0) n^{-1/2} \mathbf{M}_X \mathbf{a} = \mathbf{M}_X \mathbf{Z} \mathbf{d} + \text{résidus}, \quad (12.20)$$



**Figure 12.4** Les hypothèses nulle et alternative, le DGP, et l'angle  $\phi$

où  $\mathbf{d}$  est un vecteur à  $r$  composantes choisi par moindres carrés de façon à ce que cette régression ait un ajustement aussi bon que possible. La limite en probabilité de la somme des carrés totaux pour cette régression est l'expression (12.19) divisée par  $\sigma_0^2$ . La limite en probabilité de la somme des carrés expliqués est le NCP (12.18). Ainsi le  $R^2$  non centré de la régression (12.20) est

$$\frac{\text{plim}(n^{-1}\mathbf{a}^\top \mathbf{M}_X \mathbf{Z}) \text{plim}(n^{-1}\mathbf{Z}^\top \mathbf{M}_X \mathbf{Z})^{-1} \text{plim}(n^{-1}\mathbf{Z}^\top \mathbf{M}_X \mathbf{a})}{\text{plim}(n^{-1}\mathbf{a}^\top \mathbf{M}_X \mathbf{a})}. \quad (12.21)$$

Comme tous les  $R^2$ , on peut l'interpréter comme le carré du cosinus d'un certain angle. Dans ce cas, c'est le carré du cosinus de la limite en probabilité de l'angle formé par le vecteur  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{a}$  et la projection de ce vecteur sur le sous-espace  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$ . La limite en probabilité de cette projection est

$$\text{plim}_{n \rightarrow \infty} \left( \alpha n^{-1/2} \mathbf{M}_X \mathbf{Z} (n^{-1} \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z})^{-1} (n^{-1} \mathbf{Z}^\top \mathbf{M}_X \mathbf{a}) \right). \quad (12.22)$$

Si nous notons  $\phi$  la limite en probabilité de l'angle entre  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{a}$  et la projection (12.22), nous voyons clairement à partir de la définition d'un cosinus que  $\cos^2 \phi$  est égal au  $R^2$  (12.21).<sup>4</sup>

Tout ceci est illustré sur la Figure 12.4, pour le cas où l'hypothèse nulle ne possède qu'un seul paramètre et où une seule contrainte est testée. Le sous-espace linéaire unidimensionnel  $\mathcal{S}(\mathbf{X}_0)$  correspond à l'hypothèse nulle, et le sous-espace linéaire bidimensionnel  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$  correspond à l'hypothèse alternative. Si l'hypothèse nulle était non linéaire, nous pourrions la représenter

<sup>4</sup> Souvenons-nous que si  $\mathbf{a}$  et  $\mathbf{b}$  sont des vecteurs arbitraires, le cosinus de l'angle entre ces vecteurs est  $(\mathbf{a}^\top \mathbf{b}) / (\|\mathbf{a}\| \|\mathbf{b}\|)$ . Dans le cas particulier où  $\mathbf{a} = \mathbf{P}\mathbf{b}$ , où  $\mathbf{P}$  est une matrice de projection, la formule se simplifie en  $\|\mathbf{P}\mathbf{b}\| / \|\mathbf{b}\|$ .

sur la figure comme une variété incurvée unidimensionnelle tangente à  $\mathcal{S}(\mathbf{X}_0)$  au point  $(\beta_0, \mathbf{0})$ . Si l'hypothèse alternative était non linéaire, nous pourrions la représenter sur la figure comme une variété incurvée bidimensionnelle tangente à  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$  au point  $(\beta_0, \mathbf{0})$ , incorporant la variété unidimensionnelle correspondant à l'hypothèse nulle. Afin d'éviter toute complication sur la figure, nous n'avons représenté aucune de ces variétés. Ainsi la figure représentée suppose implicitement que les hypothèses nulle et alternative sont des modèles de régression linéaire. Cette hypothèse, cependant, est sans aucun effet sur la géométrie en cause, parce que tout dépend d'approximations linéaires quoi qu'il en soit.

Nous avons noté  $\mathbf{a}$  le DGP sur la figure. Bien sûr, le DGP est en réalité  $\mathbf{x}(\beta_0) + \alpha n^{-1/2} \mathbf{a}$ , mais nous pouvons traiter  $\mathbf{x}(\beta_0)$  comme l'origine, et puisque le facteur  $\alpha n^{-1/2}$  n'intervient pas dans les considérations géométriques, nous le fixons arbitrairement à 1 pour l'instant. L'aspect important du DGP sur la figure est qu'il n'appartient pas à l'hypothèse alternative  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$ . Il pourrait lui appartenir, bien sûr, mais comme la figure l'illustre, cela serait un cas particulier. Sur la figure, nous projetons tout d'abord  $\mathbf{a}$  sur  $\mathcal{S}^\perp(\mathbf{X}_0)$ , ce qui nous donne le point  $\mathbf{M}_X \mathbf{a}$ . Bien que  $\mathbf{a}$  corresponde à la différence entre l'hypothèse nulle simple  $\mathbf{x}(\beta_0)$  et le DGP, c'est véritablement  $\mathbf{M}_X \mathbf{a}$  qui est important pour le test, parce que c'est la différence entre  $\mathbf{a}$  et le point le plus proche appartenant à  $\mathcal{S}(\mathbf{X}_0)$  (qui est bien sûr  $\mathbf{P}_X \mathbf{a}$ ). Sur la figure, nous projetons ensuite  $\mathbf{M}_X \mathbf{a}$  sur  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$ . Tout ceci est équivalent à l'exécution de la régression (12.20). Le carré du cosinus de l'angle  $\phi$  entre  $\mathbf{M}_X \mathbf{a}$  et sa projection sur  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$  est alors l'équivalent en échantillon fini de l'expression (12.21).

La raison pour laquelle nous parlons de tests *orientés régression* devrait maintenant être claire. Si  $\mathbf{x}(\beta_0)$  est l'origine, tout modèle correspond à une direction ou ensemble de directions. L'hypothèse nulle correspond à toutes les directions dans lesquelles on peut s'éloigner de  $\mathbf{x}(\beta_0)$  tout en restant dans  $\mathcal{S}(\mathbf{X}_0)$ . Dans la Figure 12.4 il n'y a que deux directions, parce que  $\mathcal{S}(\mathbf{X}_0)$  est unidimensionnel, mais cela est un cas particulier. De manière similaire, l'hypothèse alternative correspond à toutes les directions dans lesquelles on peut s'éloigner de  $\mathbf{x}(\beta_0)$  tout en restant dans le sous-espace  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$ . Enfin, le DGP correspond à l'unique direction donnée par le vecteur  $\mathbf{a}$ . L'ensemble des directions de régression possibles est composé de toutes les directions de  $E^n$ . C'est, localement, l'ensemble de tous les DGP possibles qui laissent inchangée la structure de régression du modèle.

Revenons à l'aspect algébrique du problème. Les résultats précédents nous permettent de récrire de NCP (12.18) comme

$$\sigma_0^{-2} \alpha^2 \operatorname{plim}_{n \rightarrow \infty} (n^{-1} \mathbf{a}^\top \mathbf{M}_X \mathbf{a}) \cos^2 \phi. \quad (12.23)$$

Nous avons déjà vu que, pour un nombre donné de degrés de liberté  $r$ , la puissance asymptotique de la statistique de test (12.04) ne dépendra que de

ce NCP. Ainsi l'expression (12.23) nous enseigne tout ce qu'il est bon de connaître sur ce qui détermine la puissance asymptotique des tests orientés régression.

Le NCP (12.23) est le produit de deux facteurs. Le premier pourrait être écrit comme

$$\frac{\alpha^2 \text{plim}(n^{-1} \mathbf{a}^\top \mathbf{M}_X \mathbf{a})}{\sigma_0^2}. \quad (12.24)$$

Le numérateur de (12.24) est l'expression (12.19). C'est le carré de la limite en probabilité de la distance séparant le DGP (12.06) du point le plus proche sur une approximation linéaire de l'hypothèse nulle autour de l'hypothèse nulle simple (12.07). Le dénominateur est la variance des innovations  $\mathbf{u}$  dans le DGP (12.06), rappelant que lorsque le DGP est plus parasité, il devient plus difficile de rejeter n'importe quelle hypothèse nulle. Si nous doublons le carré de la distance entre le DGP et l'hypothèse nulle, ainsi que  $\sigma_0^2$ , le rapport (12.24) reste constant, ce qui indique que notre capacité à détecter l'inexactitude de l'hypothèse nulle reste identique. Le résultat crucial de ce rapport est qu'il ne dépend en aucun cas de  $\mathbf{Z}$ . Il sera identique pour tous les tests orientés régression de n'importe quelle hypothèse avec n'importe quel ensemble de données.

Le facteur le plus intéressant dans l'expression (12.23) est le second,  $\cos^2 \phi$ . Ce n'est qu'à travers ce facteur que le choix de  $\mathbf{Z}$  influence le NCP. Un test aura une puissance maximale, pour un nombre de degrés de liberté donné, lorsque  $\cos^2 \phi$  est égal à 1, c'est-à-dire lorsque la régression artificielle (12.20) a un  $R^2$  asymptotique égal à 1. Cela sera le cas chaque fois que le vecteur  $\mathbf{a}$  appartient au sous-espace  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_0)$  mais pas à l'espace  $\mathcal{S}(\mathbf{X}_0)$ . Autrement dit, cela sera le cas chaque fois que le DGP est un cas particulier de l'hypothèse alternative contre laquelle le test est mené, mais ne satisfait pas l'hypothèse nulle.

Par ailleurs, un test aura une puissance égale à son niveau (et par conséquent aucune **puissance utile**) lorsque  $\cos^2 \phi$  est nul. Cela surviendra lorsque  $\mathbf{a}$  appartient à  $\mathcal{S}(\mathbf{X}_0)$ , ce qui signifie que l'hypothèse nulle (ou au moins une approximation linéaire de celle-ci) est exacte. Cela surviendra également lorsque  $\mathbf{M}_X \mathbf{a}$  est asymptotiquement orthogonal à  $\mathbf{M}_X \mathbf{Z}$ , ce qui, en général, peut paraître grandement improbable. Cependant, certaines caractéristiques spéciales du modèle, ou de l'échantillon, rendent une telle situation moins rare que ce que l'on pourrait imaginer. Quoi qu'il en soit, il est sans doute peu trompeur d'affirmer que, lorsque l'hypothèse nulle est inexacte dans une direction de régression, on peut attendre de la plupart des tests orientés régression qu'ils aient une *certaine* puissance, aussi faible fût-elle.

Lorsque  $\cos^2 \phi$  est égal à 1, le NCP (12.23) est simplement

$$\frac{\alpha^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{a}^\top \mathbf{M}_X \mathbf{a} \right). \quad (12.25)$$

Puisque  $\cos^2\phi = 1$  implique que  $\mathbf{M}_X\mathbf{a}$  appartient à  $\mathcal{S}(\mathbf{M}_X\mathbf{Z})$ , cette expression peut également s'écrire

$$\frac{\alpha^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{d}^\top \mathbf{Z}^\top \mathbf{M}_X \mathbf{Z} \mathbf{d} \right) \quad (12.26)$$

pour un quelconque vecteur  $\mathbf{d}$ . Dans une analyse conventionnelle de la puissance basée sur des suites d'alternatives locales—par exemple Engle (1984)—l'hypothèse nulle serait  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}, \mathbf{0}) + \mathbf{u}$ , l'hypothèse alternative serait  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{u}$ , et le DGP serait  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0, \alpha n^{-1/2} \boldsymbol{\gamma}_0) + \mathbf{u}$ . Alors  $\mathbf{Z}$  serait la matrice  $\mathbf{X}_{\boldsymbol{\gamma}}$ , avec un élément type  $\partial x_t(\boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial \gamma_j$ , évaluée en  $(\boldsymbol{\beta}_0, \mathbf{0})$ , et  $\mathbf{d}$  serait le vecteur  $\boldsymbol{\gamma}_0$ . Le NCP (12.23) serait alors

$$\frac{\alpha^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \boldsymbol{\gamma}_0^\top \mathbf{X}_{\boldsymbol{\gamma}}^\top \mathbf{M}_X \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\gamma}_0 \right).$$

L'analyse conventionnelle est ainsi un cas particulier de l'analyse basée sur la dérive de DGP.

Les résultats précédents nous permettent de définir deux concepts neufs, qui sont quelquefois utiles dans la réflexion sur les tests. **L'hypothèse alternative implicite** d'un test est l'ensemble des DGP (c'est-à-dire le modèle ou l'ensemble des modèles) pour lequel le test a un  $\cos^2\phi$  égal à l'unité. Localement, cet ensemble doit être de dimension  $k + r$ , c'est-à-dire la dimension de l'hypothèse nulle plus le nombre de degrés de liberté de la statistique de test. Remarquons que cela pourrait comprendre davantage que **l'hypothèse alternative explicite** contre laquelle le test est mené, parce qu'il peut exister un grand nombre de modèles qui sont localement équivalents au voisinage de l'hypothèse nulle; voir Godfrey (1981) et Godfrey et Wickens (1982). A titre d'exemple, nous avons vu dans la Section 10.8 qu'une GNR, pour laquelle le régresseur de test est un vecteur de résidus retardés une fois, peut être employée contre l'hypothèse qu'un modèle de régression a des aléas AR(1) ou MA(1). Etant donné que l'une ou l'autre de ces hypothèses en tant qu'hypothèse nulle conduira exactement au même test, elles doivent appartenir toutes deux à l'hypothèse alternative implicite de ce test.

A contrario, **l'hypothèse nulle implicite** d'un test est l'ensemble des DGP contre lequel ce test aura un  $\cos^2\phi$  nul et n'aura par conséquent aucune puissance utile asymptotiquement. L'hypothèse nulle implicite doit comprendre l'hypothèse nulle de travail mais peut aussi comprendre d'autres DGP, puisque  $\cos^2\phi$  ne sera nul que si  $\mathbf{a}^\top \mathbf{M}_X \mathbf{Z} = \mathbf{0}$ . Dans certains cas, cela peut se révéler être une caractéristique souhaitable d'un test si une hypothèse implicite est large, parce qu'alors le test n'aura de la puissance que dans certaines directions. Dans d'autres cas, cependant, nous voulons que les tests soient puissants dans plusieurs directions et nous souhaiterions que l'hypothèse nulle implicite soit aussi étroite que possible.



Ces résultats montrent clairement qu'il existe un équilibre lorsque nous choisissons la direction de régression contre laquelle nous menons le test. D'un côté, nous pouvons choisir de tester contre une hypothèse alternative très contraignante, à l'aide d'un test qui ne possède qu'un seul degré de liberté. À l'opposé, nous pouvons choisir de tester contre une hypothèse alternative tout à fait générale, à l'aide d'un test à plusieurs degrés de liberté. L'accroissement du nombre de colonnes de  $\mathbf{Z}$  nous permet toujours d'augmenter  $\cos^2\phi$ , ou au pire de le laisser inchangé, ce qui augmentera la puissance de ce test. Mais ce faisant, nous augmentons  $r$ , le nombre de degrés de liberté, ce qui réduit la puissance du test. Ainsi la puissance peut soit augmenter soit diminuer lorsque nous accroissons le nombre des directions avec lesquelles nous travaillons. Cet arbitrage est au cœur d'un nombre de controverses dans la littérature consacrée aux tests d'hypothèses.

Considérons la puissance relative d'un test pour aléas AR(1) et d'un test pour aléas AR( $p$ ). Le premier ne possède qu'un seul degré de liberté, alors que le second en possède  $p$ . Le test contre des erreurs AR(1) a donc une hypothèse alternative implicite plus étroite (c'est-à-dire une hypothèse de dimension plus faible) et une hypothèse nulle implicite plus large que le test contre des erreurs AR( $p$ ). Si les aléas obéissent véritablement à un processus AR(1), il est optimal de tester contre des aléas AR(1), parce qu'un tel test aurait  $r = 1$  et  $\cos^2\phi = 1$ . Le test contre des aléas AR( $p$ ) aurait également  $\cos^2\phi = 1$  dans ce cas, mais il serait moins puissant que le test contre des aléas AR(1) parce que  $p > 1$ . Si les erreurs étaient générées par un processus AR d'ordre supérieur à 1 mais au plus égal à  $p$ , la situation serait relativement différente. À présent,  $\cos^2\phi$  serait inférieur à 1 pour le test contre des aléas AR(1), mais égal à 1 pour le test d'aléas AR( $p$ ). La différence entre les degrés de liberté pourrait encore rendre le premier test plus puissant que le second dans certains cas. Dans d'autres cas, cependant, le DGP appartiendrait véritablement à l'hypothèse nulle implicite de test d'aléas AR(1), et le second test aurait donc un niveau égal à sa puissance, asymptotiquement.

La discussion du paragraphe précédent s'applique presque sans modification à de nombreuses circonstances différentes. Par exemple, il y a eu une certaine controverse dans la littérature sur les mérites relatifs des tests d'hypothèses non emboîtées à degré de liberté unique et des tests d'englobement à degrés de liberté multiples, dont chacun a été discuté dans la Section 11.3; voir Dastoor (1983) et Mizon et Richard (1986). Les tests non emboîtés sont analogues aux tests d'aléas AR(1), les tests d'englobement sont analogues aux tests d'aléas AR( $p$ ). Nous voyons immédiatement que les tests non emboîtés doivent avoir une hypothèse alternative implicite plus petite et une hypothèse nulle implicite plus large que les tests d'englobement. Ces premiers tests seront plus puissants que les seconds si les données étaient véritablement générées par l'hypothèse non emboîtée contre laquelle le test est élaboré, mais peuvent être plus ou moins puissants dans d'autres cas.

Si nous nous écartons provisoirement de notre hypothèse de dérive de DGP et supposons que les résultats qui précèdent restent valides, nous voyons que l'arbitrage entre  $\cos^2\phi$  et les degrés de liberté est influencé par la taille de l'échantillon. Si  $n$  augmente parce que l'expérimentateur dispose de davantage d'informations, on s'attend à ce que le NCP augmente, puisqu'alors le DGP ne dérive pas vers l'hypothèse nulle lorsque la taille de l'échantillon augmente. Ainsi, on peut attendre d'une modification de  $\cos^2\phi$  un effet d'autant plus important sur la puissance que  $n$  est grand. D'autre part, l'effet de  $r$  sur la valeur critique pour le test est indépendant de la taille de l'échantillon. Ainsi, lorsque  $n$  est faible, il est particulièrement important d'employer des tests avec un nombre de degrés de liberté faible, alors que lorsque  $n$  est élevé, il est envisageable d'explorer plusieurs directions de façon à maximiser  $\cos^2\phi$ .

A proprement parler, l'analyse qui précède est incorrecte, puisque l'abandon de l'outil qu'est la dérive de DGP rend caducs les résultats sur lesquels elle se base. Cependant, une analyse Monte Carlo suggère habituellement que ces résultats correspondent assez bien en tant qu'approximations pour un DGP fixé et une taille d'échantillon fixée, à condition que le DGP soit suffisamment proche de l'hypothèse nulle et que  $n$  soit suffisamment important.<sup>5</sup> Si on les traite comme des approximations, alors on peut raisonnablement se demander ce qu'il advient lorsque  $n$  varie alors que le DGP reste fixe.

Si nous étions sûrs que l'hypothèse nulle était fausse dans une seule direction (c'est-à-dire si nous savions exactement ce que serait le vecteur  $\mathbf{a}$ ), la procédure optimale serait de n'avoir qu'une seule colonne dans  $\mathbf{Z}$ , cette colonne étant proportionnelle à  $\mathbf{a}$ . Dans la pratique, nous sommes rarement dans cette position avantageuse. Nous repérons habituellement un grand nombre d'éléments que nous supposons faux dans notre modèle et par conséquent un grand nombre de directions de régression à tester. Face à cette situation, il existe deux façons de procéder.

La première consiste à tester contre chaque type de mauvaise spécification potentielle de façon séparée, avec des tests à un ou plusieurs degrés de liberté. Si le modèle est faux dans une ou plusieurs directions de régression, cette procédure a autant de chances de nous prévenir que n'importe quelle autre. Cependant, l'expérimentateur doit rester prudent et contrôler le niveau global du test, puisque si l'on réalise, par exemple, 10 tests différents au niveau 0.05, le niveau global s'élèverait à 0.40; voir Savin (1980). De plus, il faudrait éviter de conclure trop vite que le modèle est faux sur un point particulier, simplement parce qu'une certaine statistique de test est significative. Il faut garder à l'esprit que  $\cos^2\phi$  sera souvent bien supérieur à zéro pour de *nombreux* tests, même si un seul élément est faux dans le modèle.

<sup>5</sup> Voir, par exemple, Davidson et MacKinnon (1985c). Le cas qu'ils examinent n'était pas véritablement un test orienté régression, mais comme nous le verrons dans le Chapitre 13, la théorie de la puissance des tests en général est très comparable à la théorie de la puissance des tests orientés régression.

De façon alternative, il est possible de tester un grand nombre de mauvaises spécifications simultanément en augmentant la matrice  $\mathbf{Z}$  de toutes les directions de régression que nous désirons tester. Cela maximise  $\cos^2\phi$  et par conséquent maximise l'opportunité d'obtenir un test convergent, et cela facilite le contrôle du niveau du test. Mais du fait que ce test aura de nombreux degrés de liberté, la puissance peut être faible, sauf si la taille de l'échantillon est élevée. De plus, si un tel test rejette l'hypothèse nulle, ce rejet nous procure peu d'information sur la nature de ce qui est faux dans le modèle. Bien sûr, les coefficients des colonnes individuelles de  $\mathbf{Z}$  dans la régression de test peuvent fournir de l'information.

Cela soulève le problème de ce qu'il faut faire lorsqu'un ou plusieurs tests rejettent l'hypothèse nulle. Il s'agit d'une question très difficile, et nous en discuterons dans la Section 12.7.

## 12.6 EFFICACITÉ ASYMPTOTIQUE RELATIVE

Puisque tous les tests convergents rejettent l'hypothèse nulle avec une probabilité unitaire lorsque la taille de l'échantillon tend vers l'infini, il n'est pas évident de comparer la puissance des tests dont nous ne connaissons pas les distributions asymptotiques. Des approches variées ont été proposées dans la littérature statistique, et celle qui est la plus connue est sans doute celle qui repose sur le concept de l'**efficacité asymptotique relative** ou **ARE**. Ce concept, qui est étroitement relié à l'idée d'alternatives locales, est dû à Pitman (1949), et a été développé depuis par de nombreux auteurs; consulter Kendall et Stuart (1979, Chapitre 25). Supposons que nous disposions de deux statistiques de test, disons  $\tau_1$  et  $\tau_2$ , dont les distributions asymptotiques sont identiques, et toutes deux, comme toutes les statistiques de test abordées dans ce chapitre, convergentes au taux  $n^{-1/2}$ . Cela signifie que, pour que le test ait une distribution asymptotique non dégénérée, la dérive de DGP doit approcher l'hypothèse nulle simple à un taux proportionnel à  $n^{-1/2}$ . Dans ce cas, l'efficacité asymptotique de  $\tau_2$  relativement à  $\tau_1$  est définie par

$$\text{ARE}_{21} = \lim_{n \rightarrow \infty} \left( \frac{n_1}{n_2} \right),$$

où  $n_1$  et  $n_2$  sont les tailles d'échantillon telles que  $\tau_1$  et  $\tau_2$  ont une puissance identique, et la limite est calculée lorsqu'à la fois  $n_1$  et  $n_2$  tendent vers l'infini. Si, par exemple,  $\text{ARE}_{21}$  était égale à 0.25,  $\tau_2$  nécessiterait asymptotiquement 4 fois plus d'observations que  $\tau_1$  pour atteindre la même puissance.

Pour des tests qui ont un même nombre de degrés de liberté, on voit aisément que

$$\text{ARE}_{21} = \frac{\cos^2\phi_2}{\cos^2\phi_1}.$$

**Tableau 12.1** ARE d'Autres Tests contre le Test Optimal

$r$	$\cos^2\phi$ :	1.0	0.8	0.5	0.2
1		1.000	0.800	0.500	0.200
		1.000	0.800	0.500	0.200
2		0.830	0.664	0.415	0.166
		0.775	0.620	0.388	0.155
5		0.638	0.510	0.319	0.128
		0.549	0.440	0.275	0.110
10		0.512	0.409	0.256	0.102
		0.418	0.334	0.209	0.084
20		0.402	0.322	0.201	0.080
		0.313	0.251	0.157	0.063
50		0.283	0.227	0.142	0.057
		0.210	0.168	0.105	0.042

Souvenons-nous à partir de (12.23) que le NCP est proportionnel à  $\cos^2 \phi$ . Si le DGP ne dérivait pas, il serait aussi proportionnel à la taille de l'échantillon. Si nous voulons que  $\tau_1$  et  $\tau_2$  soient de puissances identiques dans ce cas, elles doivent avoir le même NCP. Cela signifie que  $n_1/n_2$  doit être égal à l'inverse de  $\cos^2 \phi_2 / \cos^2 \phi_1$ . Supposons, par exemple, que  $\cos^2 \phi_1 = 1$  et  $\cos^2 \phi_2 = 0.5$ . Alors l'hypothèse alternative implicite pour  $\tau_1$  doit comprendre le DGP, alors que ce n'est pas le cas pour l'hypothèse alternative implicite pour  $\tau_2$ . Ainsi les directions de test de  $\tau_1$  expliquent toutes les divergences entre l'hypothèse nulle et le DGP, alors que celle de  $\tau_2$  n'en expliquent que la moitié. Mais nous pouvons compenser ce pouvoir explicatif réduit en choisissant  $n_2$  deux fois plus important que  $n_1$ , de manière à rendre les deux tests de puissances identiques asymptotiquement. Ainsi ARE<sub>21</sub> doit être égal à 0.5. Voir Davidson et MacKinnon (1987) pour davantage de détails sur ce point.

Dans le cas plus général où  $\tau_1$  et  $\tau_2$  possèdent des degrés de liberté différents, le calcul de ARE devient plus difficile. Le test optimal sera un test pour lequel l'hypothèse alternative implicite comprend la dérive de DGP (de sorte que  $\cos^2 \phi = 1$ ) et cela implique qu'il ne doit y avoir qu'un seul degré de liberté. Il peut, bien évidemment, exister un grand nombre de tests asymptotiques équivalents satisfaisant ce critère, mais il peut aussi ne pas en exister du tout dans la pratique. Les tests qui impliquent plus d'un degré de liberté, ou tels que  $\cos^2 \phi < 1$ , seront asymptotiquement moins efficaces que le test optimal et posséderont par conséquent des ARE inférieures à 1.

Les conséquences de l'usage de tests avec  $r > 1$  et/ou  $\cos^2 \phi < 1$  sont illustrées dans le Tableau 12.1. L'effet d'une modification de  $\cos^2 \phi$  ne dépend ni du niveau ni de la puissance du test, mais l'effet d'une modification de  $r$  dépend de ces deux paramètres; voir Rothe (1981) et Saikkonen (1989). Le tableau a été élaboré pour un niveau de 0.05 et des puissances de 0.90 (la

première donnée de chaque colonne) et 0.05 (la seconde donnée de chaque colonne). Chaque composante du tableau est l'ARE pour le test relativement au test optimal. Ainsi on peut interpréter chaque composante comme le facteur de proportionnalité entre la taille d'échantillon du test optimal et celle de l'autre test si tous deux doivent avoir une puissance identique asymptotiquement.

Du Tableau 12.1, nous voyons que le coût d'usage d'un test dont le nombre de degré de liberté est inutilement élevé, ou avec un  $\cos^2\phi$  de valeur inférieure à 1, peut être modique dans certains cas comme très élevé dans d'autres. Dans le pire des cas examinés, où le test non optimal est caractérisé par  $r = 50$  et  $\cos^2\phi = 0.2$ , le test optimal est tellement plus puissant que l'autre qu'il faudrait disposer d'un échantillon au moins 20 fois plus important pour le test non optimal.

## 12.7 INTERPRÉTER LE REJET DE L'HYPOTHÈSE NULLE

Supposons que l'on teste un modèle de régression dans une ou plusieurs directions et que l'on obtienne une statistique de test qui rejette l'hypothèse nulle quel que soit le niveau de signification retenu. Comment devons-nous l'interpréter? Nous avons décidé que le DGP n'appartient pas à l'hypothèse nulle implicite du test, puisque nous avons rejeté l'hypothèse nulle et donc rejeté l'hypothèse que  $\cos^2\phi$  est nul. Alors le DGP appartient-il à l'hypothèse alternative implicite? Cela est possible, mais en aucun cas obligatoire. Le NCP est le produit de l'expression (12.24), qui ne dépend pas du tout de l'hypothèse alternative du test, et de  $\cos^2\phi$ , qui lui en dépend. Pour une valeur donnée de (12.24), le NCP sera maximum lorsque  $\cos^2\phi = 1$ . Mais le fait que le NCP soit non nul (ce qui est la seule information livrée par la statistique de test) implique seulement que ni  $\cos^2\phi$  ni (12.24) n'est nul. Ainsi la seule conclusion que nous puissions tirer d'une seule statistique de test significative est que le DGP n'est pas un cas particulier du modèle soumis au test et que les directions représentées par  $\mathbf{Z}$  ont un certain pouvoir explicatif pour la direction  $\mathbf{a}$  dans laquelle le modèle est véritablement inexact.

Si nous voulons faire une quelconque inférence sur les directions dans lesquelles le modèle soumis au test est faux, nous devons à l'évidence calculer plus d'une statistique de test. Puisque l'expression (12.24) est identique pour tous les tests orientés régression, toutes les différences entre les valeurs des diverses statistiques de test doivent provenir de différences entre les nombres de degrés de liberté, entre les  $\cos^2\phi$ , ou tout simplement être aléatoires (et parmi elles des différences entre les comportements avec des échantillons finis et asymptotiques des tests). Supposons que l'on teste contre certains ensembles de directions de régression, représentés par les matrices  $\mathbf{Z}_1$ ,  $\mathbf{Z}_2$ , et ainsi de suite. Supposons par ailleurs que la  $j^{\text{ième}}$  matrice,  $\mathbf{Z}_j$ , possède  $r_j$  colonnes et génère une statistique de test  $T_j$ , distribuée asymptotiquement suivant une  $\chi^2(r_j)$  sous l'hypothèse nulle. On peut employer chacune des statistiques  $T_j$

pour estimer le NCP correspondant, disons  $\Lambda_j$ . Puisque l'espérance d'une variable aléatoire de la distribution  $\chi^2$  non centrée à  $r$  degrés de liberté est la somme de  $r$  et du NCP, l'estimation évidente de  $\Lambda_j$  est  $T_j - r_j$ . Evidemment, cet estimateur n'est pas convergent, puisque sous une dérive de DGP la statistique de test est une variable aléatoire quelle que soit la taille de l'échantillon. Quoi qu'il en soit, si  $T_l - r_l$  est sensiblement inférieure à  $T_j - r_j$  pour tout  $j \neq l$ , on peut logiquement rechercher un meilleur modèle dans les directions testées par  $\mathbf{Z}_l$ .

Il n'est pas du tout certain que  $\mathbf{Z}_l$ , la matrice de régresseurs avec le NCP estimé le plus élevé, représente vraiment les directions omises. Après tout, il est fort possible que nous ne testions pas du tout les bonnes directions, auquel cas  $\mathbf{M}_X \mathbf{a}$  peut ne pas appartenir au sous-espace  $\mathcal{S}(\mathbf{X}_0, \mathbf{Z}_j)$  quel que soit  $j$ . Cependant, la modification du modèle dans les directions représentées par  $\mathbf{Z}_l$  sera une stratégie raisonnable dans bien des cas, en particulier lorsque  $\mathbf{Z}_l$  possède peu de colonnes et que  $T_l - r_l$  est sensiblement supérieure aux autres NCP estimés. Une attitude possible consiste à construire une matrice de régresseurs de test  $\mathbf{Z}_J$  telle qu'elle engendre le sous-espace engendré par toutes les  $\mathbf{Z}_j$ . Autrement dit,  $\mathbf{Z}_J$  doit être la "réunion" de toutes les colonnes des  $\mathbf{Z}_j$ . Ainsi la statistique de test  $T_J$  correspondant à  $\mathbf{Z}_J$  doit être supérieure à n'importe quelle autre statistique de test. Dans ce cas, si  $T_J$  était à peine supérieure à  $T_l$ , et en particulier si elle n'en était pas supérieure de plus que la différence entre les degrés de liberté, on pourrait penser à raison que les directions représentées par  $\mathbf{Z}_l$  rendent compte de façon satisfaisante des différences entre l'hypothèse nulle et le DGP.

L'examen d'un exemple simple et fréquent peut aider à fixer les idées développées jusqu'à présent. Supposons que l'hypothèse nulle soit

$$H_0: y_t = \mathbf{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2),$$

où  $\mathbf{X}_t$  est un vecteur ligne, et que nous nous intéressions à la tester contre deux hypothèses alternatives distinctes,

$$H_1: y_t = \mathbf{X}_t \boldsymbol{\beta} + \rho(y_{t-1} - \mathbf{X}_{t-1} \boldsymbol{\beta}) + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \text{ et } \quad (12.27)$$

$$H_2: y_t = \mathbf{X}_t \boldsymbol{\beta} + \delta y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \quad (12.28)$$

Ainsi l'hypothèse nulle  $H_0$  est emboîtée à la fois dans  $H_1$  et  $H_2$ . La première alternative modifie  $H_0$  en lui associant des aléas AR(1) alors que la seconde la modifie en lui associant la variable dépendante retardée.

Notre but est de calculer les NCP et les valeurs correspondantes de  $\cos^2 \phi$  pour les tests de  $H_0$  contre  $H_1$  et  $H_2$  lorsque les données sont générées par (12.28). Ainsi nous supposerons que les données sont générées par une dérive de DGP qui est un cas particulier de  $H_2$ . Cette dérive peut s'écrire comme

$$y_t = \mathbf{X}_t \boldsymbol{\beta}_0 + \alpha_0 n^{-1/2} (\mathbf{X}_{t-1} \boldsymbol{\beta}_0 + u_{t-1}) + u_t, \quad u_t \sim \text{IID}(0, \sigma_0^2). \quad (12.29)$$

Notons que ce DGP n'implique pas le calcul récursif de  $y_t$ , contrairement à (12.28), parce que (12.29) est localement équivalente à (12.28) au voisinage de  $\delta = 0$  et  $\alpha_0 = 0$ .

Lorsque nous testons  $H_0$  contre  $H_2$ , nous testerons dans la direction du DGP et  $\cos^2\phi$  sera bien sûr égal à 1. À l'aide de l'expression (12.25), nous voyons que le NCP pour ce test est

$$A_{22} \equiv \frac{\alpha_0^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1})^\top \mathbf{M}_X (\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1}) \right), \quad (12.30)$$

où  $\mathbf{u}_{-1}$  et  $\mathbf{X}_{-1}$  désignent respectivement le vecteur dont l'élément type est  $u_{t-1}$  et la matrice dont la ligne type est  $\mathbf{X}_{t-1}$ . Ici,  $\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1}$  joue le rôle du vecteur  $\mathbf{a}$  dans l'expression (12.25). La notation  $A_{22}$  signifie que  $H_2$  est l'alternative contre laquelle le test est mené et que le DGP appartient à  $H_2$ . Le calcul de la limite en probabilité donne

$$\begin{aligned} A_{22} &= \frac{\alpha_0^2}{\sigma_0^2} \left( \sigma_0^2 + \text{plim}_{n \rightarrow \infty} \frac{1}{n} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2 \right) \\ &= \alpha_0^2 \left( 1 + \sigma_0^{-2} \text{plim}_{n \rightarrow \infty} \frac{1}{n} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2 \right). \end{aligned}$$

Examinons à présent ce qu'il advient lorsque nous testons  $H_0$  contre  $H_1$ . Au voisinage de  $H_0$ , cette dernière est localement équivalente à

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho \mathbf{u}_{-1} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (12.31)$$

ce qui évite le calcul récursif que (12.27) semble impliquer. Parce que les processus AR(1) et MA(1) sont localement équivalents aux alentours du point où leurs paramètres respectifs sont nuls, cela ressemble à un processus à erreurs MA(1). Nous voyons à partir de (12.31) que  $\mathbf{u}_{-1}$  remplace de  $\mathbf{Z}$ . Comme auparavant,  $\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1}$  remplace de  $\mathbf{a}$ . Ainsi, à partir de (12.28), le NCP est donné par

$$\begin{aligned} A_{12} &= \frac{\alpha_0^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1})^\top \mathbf{M}_X \mathbf{u}_{-1} \right) \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{u}_{-1}^\top \mathbf{M}_X \mathbf{u}_{-1} \right)^{-1} \\ &\quad \times \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{u}_{-1}^\top \mathbf{M}_X (\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1}) \right). \end{aligned} \quad (12.32)$$

Parce que

$$\begin{aligned} &\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{X}_{-1}\boldsymbol{\beta}_0 + \mathbf{u}_{-1})^\top \mathbf{M}_X \mathbf{u}_{-1} \right) \\ &= \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\boldsymbol{\beta}_0^\top \mathbf{X}_{-1}^\top \mathbf{M}_X \mathbf{u}_{-1} + \mathbf{u}_{-1}^\top \mathbf{M}_X \mathbf{u}_{-1}) \right) = \sigma_0^2, \end{aligned}$$

l'expression (12.32) se simplifie en

$$\frac{\alpha_0^2}{\sigma_0^2} \sigma_0^2 (\sigma_0^{-2}) \sigma_0^2 = \alpha_0^2.$$

Comme les données ont été générées par un cas particulier de  $H_2$ ,  $\cos^2 \phi$  pour le test contre  $H_1$  est simplement le rapport du NCP  $A_{12}$  au NCP  $A_{22}$ . Ainsi

$$\begin{aligned} \cos^2 \phi &= \alpha_0^2 \left( \alpha_0^2 \left( 1 + \sigma_0^{-2} \text{plim } \frac{1}{n} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2 \right) \right)^{-1} \\ &= \left( 1 + \frac{\text{plim } n^{-1} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2}{\sigma_0^2} \right)^{-1}. \end{aligned} \quad (12.33)$$

La seconde ligne de (12.33) fournit une expression remarquablement simple pour  $\cos^2 \phi$  dans ce cas spécial. Il ne dépend que du rapport de la limite en probabilité de  $n^{-1}$  fois la norme au carré du vecteur  $\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0$  à la variance des aléas dans le DGP (12.29). Lorsque ce rapport tend vers zéro,  $\cos^2 \phi$  tend vers un. À l'opposé, lorsque ce rapport tend vers l'infini,  $\cos^2 \phi$  tend vers zéro. L'intuition est assez simple. Lorsque le rapport de  $\text{plim } n^{-1} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2$  à  $\sigma_0^2$  tend vers zéro, parce que par exemple  $\boldsymbol{\beta}_0$  tend vers zéro,  $\mathbf{M}_X \mathbf{y}_{-1}$  (où  $\mathbf{y}_{-1}$  est d'élément type  $y_{t-1}$ ) se confond avec  $\mathbf{M}_X \mathbf{u}_{-1}$ . Lorsque c'est le cas, un test contre  $H_1$  se confond avec un test contre  $H_2$ . D'autre part, lorsque le rapport tend vers l'infini, la corrélation entre  $y_{t-1}$  et  $u_{t-1}$  tend vers zéro et les directions pour lesquelles  $H_1$  et  $H_2$  divergent de  $H_0$  tendent à être mutuellement orthogonales.

L'analyse que nous venons de mener s'applique aussi aisément sous l'hypothèse que les colonnes ont été générées par un cas particulier de  $H_1$ . La dérive de DGP serait alors

$$y_t = \mathbf{X}_t \boldsymbol{\beta}_0 + \rho_0 n^{-1/2} u_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma_0^2).$$

Lorsque nous testons  $H_0$  contre  $H_1$ ,  $\cos^2 \phi$  est égal à 1, et par un argument encore plus simple que celui qui nous a conduit à (12.32) nous voyons que le NCP est

$$A_{11} = \frac{\rho_0^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{u}_{-1}^\top \mathbf{M}_X \mathbf{u}_{-1} \right) = \rho_0^2.$$

De manière comparable, lorsque nous testons  $H_0$  contre  $H_2$ , le NCP est

$$\begin{aligned} A_{21} &= \frac{\rho_0^2}{\sigma_0^2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{u}_{-1}^\top \mathbf{M}_X (\mathbf{X}_{-1} \boldsymbol{\beta}_0 + \mathbf{u}_{-1}) \right) \\ &\quad \times \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{X}_{-1} \boldsymbol{\beta}_0 + \mathbf{u}_{-1})^\top \mathbf{M}_X (\mathbf{X}_{-1} \boldsymbol{\beta}_0 + \mathbf{u}_{-1}) \right)^{-1} \\ &\quad \times \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{X}_{-1} \boldsymbol{\beta}_0 + \mathbf{u}_{-1})^\top \mathbf{M}_X \mathbf{u}_{-1} \right). \end{aligned}$$



Cette expression se simplifie finalement:

$$\begin{aligned} & \frac{\rho_0^2}{\sigma_0^2} \sigma_0^2 \left( \sigma_0^2 + \text{plim } \frac{1}{n} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2 \right)^{-1} \sigma_0^2 \\ &= \rho_0^2 \left( 1 + \sigma_0^{-2} \text{plim } \frac{1}{n} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2 \right)^{-1}. \end{aligned}$$

Bien sûr,  $\cos^2 \phi$  pour le test de  $H_0$  contre  $H_2$  est l'expression du membre de droite divisé par  $\rho_0^2$ , soit

$$\left( 1 + \frac{\text{plim } n^{-1} \|\mathbf{M}_X \mathbf{X}_{-1} \boldsymbol{\beta}_0\|^2}{\sigma_0^2} \right)^{-1}. \quad (12.34)$$

Il est sans doute utile de commenter ce dernier résultat. Nous avons vu que  $\cos^2 \phi$  pour le test contre  $H_2$  lorsque les données ont été générées par  $H_1$ , l'expression (12.34), est identique à  $\cos^2 \phi$  pour le test contre  $H_1$  lorsque les données ont été générées par  $H_2$ , l'expression (12.33). Ce résultat n'est pas spécifique à cet exemple, mais reste valable chaque fois que les alternatives impliquent des tests à un seul degré de liberté. D'un point de vue géométrique, cette équivalence reflète simplement le fait que lorsque  $\mathbf{z}$  est un vecteur, l'angle formé par  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{a}$  et sa projection  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{a}$  sur  $\mathcal{S}(\mathbf{X}, \mathbf{z})$ , qui est

$$\alpha n^{-1/2} \mathbf{M}_X \mathbf{z} (\mathbf{z}^\top \mathbf{M}_X \mathbf{z})^{-1} \mathbf{z}^\top \mathbf{M}_X \mathbf{a},$$

est le même que l'angle formé par  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{a}$  et  $\alpha n^{-1/2} \mathbf{M}_X \mathbf{z}$ . Cela provient du fait que  $(\mathbf{z}^\top \mathbf{M}_X \mathbf{z})^{-1} \mathbf{z}^\top \mathbf{M}_X \mathbf{a}$  est un scalaire lorsque  $\mathbf{z}$  est un vecteur. Donc, si nous inversons les positions de  $\mathbf{a}$  et  $\mathbf{z}$ , l'angle reste inchangé. Cette propriété géométrique provient également de deux propriétés numériques. Premièrement, dans les régressions

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\alpha} + \gamma \mathbf{z} + \text{résidus} \quad \text{et} \\ \mathbf{z} &= \mathbf{X}\boldsymbol{\beta} + \delta \mathbf{y} + \text{résidus}, \end{aligned}$$

le  $t$  de Student de  $\mathbf{z}$  dans la première est égal à celui de  $\mathbf{y}$  dans la seconde. Ensuite, dans les régressions

$$\begin{aligned} \mathbf{M}_X \mathbf{y} &= \gamma \mathbf{M}_X \mathbf{z} + \text{résidus} \quad \text{et} \\ \mathbf{M}_X \mathbf{z} &= \delta \mathbf{M}_X \mathbf{y} + \text{résidus}, \end{aligned}$$

les  $t$  de Student de  $\gamma$  et  $\delta$  sont numériquement identiques ainsi que les  $R^2$  non centrés.

L'analyse de la puissance pour cet exemple illustre la simplicité et la généralité de l'idée de dérive de DGP. Bien que le cas considéré soit plutôt simple, c'est un cas fréquent dans la pratique. Les modèles de régression estimés à l'aide de données chronologiques manifestent souvent l'existence d'une autocorrélation sous la forme de statistiques de Durbin-Watson faibles ou d'autres statistiques de test pour aléas AR(1) significatives. Nous avons vu que la présence d'une telle corrélation est presque aussi compatible avec l'hypothèse que le modèle devrait comprendre une variable dépendante retardée qu'avec l'hypothèse que les aléas suivent un processus AR(1) (excepté lorsque  $\text{plim } n^{-1} \|M_X X_{-1} \beta_0\|^2$  est relativement importante par rapport à  $\sigma_0^2$ ). Ainsi il faudrait rester très prudent en interprétant les résultats d'un test contre des aléas AR(1) qui rejette l'hypothèse nulle. On voudrait sûrement envisager de nombreux modèles alternatifs en plus de l'alternative que les aléas obéissent vraiment à un processus AR(1). En dernière limite, avant même d'accepter provisoirement cette alternative, on voudrait la soumettre à des tests des contraintes du facteur commun dont nous avons discuté dans la Section 10.9.

Dans l'exemple précédent, il était facile d'évaluer de manière analytique les valeurs de  $\Lambda$  et  $\cos^2 \phi$  qui nous intéressaient. Cela ne sera pourtant pas toujours le cas. Cependant, il est toujours possible de calculer des approximations à ces quantités. Pour cela, il suffit d'exécuter la régression (12.20), en évaluant  $X(\beta)$ ,  $\mathbf{a}$ , et  $\mathbf{Z}$  avec les valeurs des paramètres supposées (ou estimées). Si  $\mathbf{a}$  et/ou  $\mathbf{Z}$  étaient stochastiques, il faudrait les générer de façon aléatoire et employer un grand nombre d'observations (que l'on peut obtenir en multipliant les observations disponibles aussi souvent que nécessaire) afin d'approximer les limites en probabilité pertinentes. Le  $R^2$  non centré de la régression fournit une approximation de  $\cos^2 \phi$  et la somme des carrés expliqués fournit une approximation de  $\Lambda$ .

## 12.8 LE NON REJET DE L'HYPOTHÈSE NULLE

Pour la grande part de ce chapitre, nous avons focalisé notre attention sur l'interprétation des statistiques de test qui rejettent l'hypothèse nulle. Dans de nombreuses circonstances, bien sûr, les statistiques de test ne la rejettent pas. Ainsi, il est tout aussi important de maîtriser l'interprétation du rejet que celle du non rejet. Bien que nous employions quelquefois le terme "acceptation" de l'hypothèse nulle lorsqu'une ou plusieurs statistiques de test ne la rejettent pas, une telle acceptation ne peut être que provisoire et doit être modulée avec précaution. L'intensité de notre précaution dépend de la puissance du (des) test(s) qui n'a (n'ont) pas rejeté l'hypothèse nulle. Nous pouvons faire davantage confiance en la validité de l'hypothèse nulle si les tests reconnus pour leur grande puissance contre les alternatives ne la rejettent pas.

Comme nous l'avons vu, la puissance d'un test dépend de la manière dont les données ont été générées. Dans un article récent, Andrews (1989) a

suggéré que, pour aider à l'interprétation du non rejet d'une hypothèse nulle par un test particulier, il faudrait considérer la puissance qu'aurait le test sous les DGP associés aux hypothèses alternatives d'intérêt. Il semble raisonnable que de telles alternatives ne soient pas écartées à la faveur de l'hypothèse nulle sur la base des tests qui auraient, sous ces alternatives, une probabilité faible de rejeter l'hypothèse nulle. Autrement dit, on ne doit pas dire qu'un test a *discriminé* contre une alternative en faveur de l'hypothèse nulle s'il aurait une chance faible de rejeter l'hypothèse nulle même si l'hypothèse alternative était exacte.

L'outil analytique employé par Andrews est la **fonction puissance inverse** qui, comme son nom l'indique, est reliée à la fonction puissance dont nous avons discuté dans la Section 12.3. Pour nos besoins immédiats, nous supposons que les hypothèses alternatives d'intérêt peuvent s'exprimer en termes d'un ensemble de paramètres et que l'hypothèse nulle correspond à un ensemble de contraintes sur ces paramètres. Alors, pour un niveau de test  $\alpha$  et pour une puissance désirée  $\pi$ , la fonction puissance inverse pour une statistique de test donnée spécifie les valeurs paramétriques qui caractérisent les DGP qui ont une puissance  $\pi$  de rejeter l'hypothèse nulle pour un test de niveau  $\alpha$ . Si les valeurs paramétriques données par la fonction puissance inverse sont proches des valeurs paramétriques issues des contraintes de l'hypothèse nulle, un non rejet de l'hypothèse nulle peut s'interpréter comme le fait que l'hypothèse nulle n'est pas véritablement fautive dans une direction quelconque correspondant aux différentes alternatives. Si, autrement, la fonction puissance inverse produit des valeurs paramétriques éloignées de l'hypothèse nulle, un non rejet nous indique peu de choses sur l'exactitude de l'hypothèse nulle, puisque ce non rejet est compatible avec de nombreuses alternatives possibles.

Andrews montre la procédure de calcul des fonctions puissance inverse pour une large classe de tests asymptotiques pour des contraintes uniques et multiples. Nous n'examinerons que le cas de la contrainte unique, parce qu'il est beaucoup plus simple que l'autre cas. Supposons que l'hypothèse d'intérêt est qu'un certain paramètre, disons  $\theta$ , prend une valeur donnée, disons  $\theta_0$ . Pour être concret, nous pourrions supposer que  $\theta$  est un paramètre d'une fonction de régression non linéaire. Il existe de nombreuses statistiques de test asymptotiquement équivalentes, parmi lesquelles la plus simple est

$$\frac{\hat{\theta} - \theta_0}{\hat{\sigma}_{\theta}}. \quad (12.35)$$

Puisque le dénominateur est ici une estimation de l'écart type de  $\hat{\theta}$ , (12.35) est simplement un  $t$  de Student asymptotique. Cette statistique de test est asymptotiquement équivalente à la racine carrée de (12.04).

En considérant (12.35), nous trahissons notre engagement de ne considérer que des statistiques asymptotiquement distribuées selon une  $\chi^2$ . Cela se justifie par les avantages de la simplicité. Considérons la dérivée de DGP

pour laquelle  $\theta = \theta_0 + n^{-1/2}\delta$ , et supposons que sous ce DGP  $\hat{\sigma}_\theta \stackrel{a}{=} n^{-1/2}\tau$ , pour un quelconque  $\tau = O(1)$  lorsque  $n \rightarrow \infty$ , puisque  $\hat{\theta}$  est convergent au taux  $n^{-1/2}$ . Alors la distribution asymptotique de (12.35) est  $N(\lambda, 1)$ , avec  $\lambda = \delta/\tau$ . Cette simple propriété nous autorise à calculer la fonction puissance asymptotique de la statistique (12.35). Si la valeur critique pour un test bilatéral de niveau  $\alpha$  basé sur la distribution  $N(0, 1)$  est désignée par  $c_\alpha$ , la probabilité de rejeter l'hypothèse nulle sous notre dérive de DGP est la probabilité qu'une variable aléatoire distribuée suivant une  $N(\lambda, 1)$  ait une valeur absolue supérieure à  $c_\alpha$ . Soit  $\Phi(\cdot)$  la c.d.f. de la distribution normale centrée et réduite, cette probabilité est

$$P(\alpha, \lambda) \equiv 1 - \Phi(c_\alpha - \lambda) + \Phi(-c_\alpha - \lambda). \quad (12.36)$$

Afin de trouver la fonction puissance inverse correspondant à (12.36), nous posons  $P(\alpha, \lambda) = \pi$  pour un niveau de puissance désiré  $\pi$ . Cette équation définit implicitement la fonction puissance inverse. Il est aisé de vérifier à partir de (12.36), que  $P(\alpha, -\lambda) = P(\alpha, \lambda)$ . Ainsi, si  $P(\alpha, \lambda) = \pi$ , alors  $P(\alpha, -\lambda) = \pi$  également. Cependant, la non unicité de  $\lambda$  disparaîtrait si nous calculions le carré de la statistique de test pour obtenir une forme  $\chi^2$ . Il n'existe aucune expression comparable donnant la valeur (absolue) de  $\lambda$  comme une fonction de  $\alpha$  et  $\pi$  dans l'exemple présent, mais pour des arguments donnés,  $\lambda$  n'est pas difficile à calculer numériquement.

Quelle interprétation donner à la fonction  $\lambda(\alpha, \pi)$ ? Si nous élevons au carré la statistique asymptotiquement normale (12.35) pour obtenir une forme  $\chi^2$ , le résultat aura une distribution limite  $\chi^2(1, \Lambda)$  avec  $\Lambda = \lambda^2$ . Alors il apparaît que  $\Lambda = (\lambda(\alpha, \pi))^2$  est asymptotiquement le NCP le plus faible nécessaire pour qu'un test de niveau  $\alpha$  basé sur le carré de (12.35) ait une probabilité de rejeter l'hypothèse nulle au moins égale à  $\pi$ .

Soit le modèle de régression non linéaire écrit sous sa forme habituelle

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad (12.37)$$

où le paramètre d'intérêt  $\theta$  est un élément du vecteur de paramètres  $\boldsymbol{\beta}$ . Si nous notons  $\mathbf{X}_\theta$  la dérivée du vecteur  $\mathbf{x}(\boldsymbol{\beta})$  par rapport à  $\theta$ , évaluée avec les paramètres  $\boldsymbol{\beta}_0$ , et  $\mathbf{M}_X$  la projection sur le complément orthogonal de l'espace engendré par toutes les colonnes de  $\mathbf{X}(\boldsymbol{\beta})$  autre que  $\mathbf{X}_\theta$ , alors la variance asymptotique de l'estimateur des moindres carrés  $\hat{\theta}$  est  $\sigma_0^2(\mathbf{X}_\theta^\top \mathbf{M}_X \mathbf{X}_\theta)^{-1}$ , où  $\sigma_0^2$  est la variance des composantes de  $\mathbf{u}$ . Si nous considérons un DGP avec un paramètre  $\theta \neq \theta_0$ , alors pour une taille d'échantillon  $n$ , le paramètre  $\delta$  de la dérive de DGP devient  $n^{1/2}(\theta - \theta_0)$ , et  $\Lambda = \lambda^2$  devient

$$\Lambda = \frac{1}{\sigma_0^2}(\theta - \theta_0)^2 \mathbf{X}_\theta^\top \mathbf{M}_X \mathbf{X}_\theta. \quad (12.38)$$

On peut comparer avec l'expression générale (12.36). Posons maintenant  $\theta(\alpha, \pi)$  comme la valeur de  $\theta$  qui égalise  $\Lambda$  dans (12.38) à  $(\lambda(\alpha, \pi))^2$ . Nous

voyons que, à l'intérieur de cette approximation asymptotique, les DGP dont les valeurs de  $\theta$  sont plus proches du  $\theta_0$  de l'hypothèse nulle que  $\theta(\alpha, \pi)$  auront une probabilité inférieure à  $\pi$  de rejeter l'hypothèse nulle sur un test de niveau  $\alpha$ .

Nous souhaiterions ne pas considérer le non rejet de l'hypothèse nulle comme une évidence contre d'autres DGP ou ensemble de DGP si, sous ces derniers, la probabilité de rejeter l'hypothèse nulle n'est pas suffisamment élevée. Qu'entendons-nous par "suffisamment élevée"? On peut exercer l'intuition à ce sujet en considérant ce que nous apprendrions sur le contexte présent en employant un outil ordinaire de l'inférence statistique conventionnelle, à savoir l'intervalle de confiance. Armés de l'estimation de  $\hat{\theta}$  et d'une estimation de son écart type,  $\hat{\sigma}_\theta$ , nous pouvons construire un intervalle de confiance  $[\hat{\theta} - c_\alpha \hat{\sigma}_\theta, \hat{\theta} + c_\alpha \hat{\sigma}_\theta]$ . Sous l'hypothèse conventionnelle que le DGP est obtenu en choisissant des valeurs spécifiques des paramètres de la régression non linéaire (12.37), cet intervalle de confiance a une probabilité proche de  $1 - \alpha$ , pour des échantillons importants, de comprendre le véritable paramètre. Aucune hypothèse nulle caractérisée par  $\theta_0$  à l'intérieur de l'intervalle de confiance ne sera rejetée par un test de niveau  $\alpha$ . Un intervalle de confiance est *aléatoire*: il dépend de la valeur réalisée de l'estimation  $\hat{\theta}$ . Au contraire, la fonction puissance inverse est déterministe, aussi devons-nous être prudents dans nos analogies. Cependant, il semble raisonnable que, lorsque nous désirons nous abstraire des ensembles de données réalisés, nous devrions refuser de considérer l'éventualité du non rejet d'une hypothèse nulle comme l'évidence contre tout DGP dont les paramètres appartiennent à la région de confiance de taille comparable à l'intervalle de confiance.

Que cela implique-t-il pour le choix de la puissance désirée  $\pi$ ? Une réponse approximative à cette question est très facile à trouver. Supposons que dans (12.38) nous réclamions que  $\theta - \theta_0$  divisée par l'écart type de  $\hat{\theta}$  soit égale à  $c_\alpha$ . Cela signifie précisément que la différence entre  $\theta$  et  $\theta_0$  est la moitié de la longueur de l'intervalle de confiance associé à un niveau  $\alpha$  pour la valeur donnée de l'écart type. Pour des paramètres  $\alpha$  et  $\pi$  donnés, la valeur de la fonction puissance inverse  $\lambda(\alpha, \pi)$  implique une valeur de  $\theta$ , selon (12.38). Nous pourrions donc nous demander quelle valeur de  $\pi$  produira la condition requise sur l'écart  $\theta - \theta_0$ . Cette valeur  $\pi$  est évidemment la solution de l'équation  $\lambda(\alpha, \pi) = c_\alpha$ , où, en termes de la fonction puissance inverse  $P$  elle-même,  $P(\alpha, c_\alpha) = \pi$ . Si désormais nous remplaçons  $P$  par son expression explicite provenant de (12.36), nous réclamons que

$$\pi = 1 - \Phi(0) + \Phi(-2c_\alpha) = \frac{1}{2} + \Phi(-2c_\alpha).$$

Pour des choix raisonnables de  $\alpha$ , le dernier terme sera extrêmement faible. Par exemple, si  $\alpha = .05$ , de sorte que  $c_\alpha \cong 1.96$ , un petit calcul nous montre que  $\Phi(-3.92) = .0000443$ . Par conséquent, avec une approximation très satisfaisante, nous obtenons  $\pi = \frac{1}{2}$ , indépendamment de  $\alpha$ .

Ce résultat est compatible avec l'intuition. En s'éloignant de la valeur de  $\theta_0$  associée à une hypothèse nulle quelconque d'une quantité qui correspond à la moitié de la longueur de l'intervalle de confiance pour tout niveau de test raisonnable, nous obtenons les valeurs des paramètres associés aux DGP qui ont une probabilité de 0.5 de rejeter l'hypothèse nulle sur un test de niveau identique.

D'autres choix de  $\pi$  sont bien sûr envisageables. Un choix qui paraît naturel dans certains contextes est  $\pi = 1 - \alpha$ , ce qui rend le risque de première espèce égal au risque de deuxième espèce dans un certain sens. Lors du choix du niveau  $\alpha$ , nous acceptons l'éventualité du rejet d'une hypothèse nulle exacte avec une probabilité  $\alpha$ . Lorsque nous refusons de traiter le non rejet d'une hypothèse nulle par un test de niveau  $\alpha$  comme l'évidence contre des valeurs paramétriques qui génèrent des NCP plus faibles que la fonction puissance inverse évaluée en  $\alpha$  et  $1 - \alpha$ , nous acceptons le fait que ces valeurs paramétriques que nous rejetons, sur la base du non rejet de l'hypothèse nulle, n'auraient pas rejeté l'hypothèse nulle avec une probabilité  $\alpha$ .

Il faut prendre d'innombrables précautions à ce stade. La totalité de l'analyse précédente se fonde sur l'hypothèse que le vrai DGP appartient à la classe des DGP que l'on peut décrire par un modèle de régression non linéaire (12.37). Il existe en général un grand nombre de DGP qui ne satisfont pas (12.37) pour lesquels la probabilité de rejeter une hypothèse nulle donnée satisfaisant (12.37) est faible. Typiquement, de tels DGP impliqueraient des variables explicatives plus nombreuses ou plus pertinentes que dans (12.37). Hélas, un rejet ou un non rejet d'une hypothèse nulle basée sur l'écriture (12.37) ne nous dit rien sur la possible existence d'un meilleur modèle. C'est du talent de l'économètre, plutôt que de procédures de test, que dépend l'élaboration de modèles potentiellement meilleurs qui seront ultérieurement soumis à des procédures de test formelles.

Bien que notre exposé théorique fût facilité par l'usage de la fonction puissance (12.36) basée sur la distribution normale, dans la pratique, lorsque l'on veut calculer des fonctions puissance inverses, il est plus aisé d'employer les propriétés de la distribution du  $\chi^2$  non centrée. Soit  $c_\alpha(r)$  la valeur critique pour un test de niveau  $\alpha$  basé sur la distribution du  $\chi^2$  centrée à  $r$  degrés de liberté. Alors la probabilité qu'une variable aléatoire suivant la distribution  $\chi^2(r, \Lambda)$  prenne une valeur supérieure à  $c_\alpha(r)$  peut s'exprimer en termes de la c.d.f.  $F_{(r, \Lambda)}(\cdot)$  de cette distribution. La probabilité adéquate est simplement  $1 - F_{(r, \Lambda)}(c_\alpha(r))$ . Par conséquent, la fonction puissance inverse s'obtient en résolvant l'équation en  $\Lambda$  en termes de  $r$ ,  $\alpha$ , et  $\pi$ :

$$\pi = 1 - F_{(r, \Lambda)}(c_\alpha(r)).$$

La valeur de  $\Lambda$  solution de cette équation peut s'utiliser dans une formule telle que (12.38) afin de déterminer les valeurs paramétriques qui ont vraiment généré les NCP égaux à  $\Lambda$ .

**Table 12.2** Quelques Valeurs de  $\Lambda(1, \alpha, \pi)$ 

$\alpha$	$\pi:$	.50	.90	.95	.99
0.10		2.701	8.564	10.822	15.770
0.05		3.841	10.507	12.995	18.372
0.01		6.635	14.879	17.814	24.031

Andrews (1989) fournit des valeurs de la fonction puissance inverse, que l'on peut noter  $\Lambda(r, \alpha, \pi)$ , pour une variété de valeurs de  $r$ ,  $\alpha$ , et  $\pi$ , mais les ordinateurs modernes et leurs logiciels rendent caduc l'usage de ces tables. Tout programme capable de calculer la c.d.f. de la distribution du  $\chi^2$  non centrée peut être utilisé également pour le calcul de la fonction puissance inverse. Afin de ne pas pénaliser les lecteurs qui n'ont pas de programme disponible pour l'instant, nous reportons des valeurs significatives dans le Tableau 12.2.

Considérons à présent un exemple simple de l'usage de la fonction puissance inverse. Supposons que  $\theta_0$  soit égal à 1 et que l'écart type de  $\hat{\theta}$  soit 0.60. Alors pour un test de niveau 0.05, les valeurs de  $\theta$  données par la fonction puissance inverse pour  $\pi = .5$  sont  $-0.176$  et  $2.176$ . Ainsi, pour tout  $\theta$  compris entre ces bornes, la probabilité que le test rejette l'hypothèse nulle est inférieure à .5. Si au lieu de cela nous choisissons  $\pi = 1 - \alpha = .95$ , les valeurs données par la fonction puissance inverse seraient  $-0.974$  et  $2.974$ , un intervalle plus large à l'intérieur duquel la probabilité que le test rejette l'hypothèse nulle est inférieure à .95.

Cet exemple illustre la manière d'employer la fonction puissance inverse. Elle offre un moyen simple de connaître les valeurs de  $\theta$  pour lesquelles le test a toutes les chances d'avoir une puissance faible ou forte. La fonction puissance inverse est extrêmement facile à calculer, du moins pour les tests de contrainte unique. Ainsi, il semble utile de la calculer chaque fois qu'un test de contrainte unique conduit ou non au rejet de l'hypothèse nulle. Les fonctions puissance inverse peuvent également être calculées pour des tests de contraintes multiples, mais les calculs sont plus difficiles et l'interprétation plus délicate. Les lecteurs devraient consulter l'article de Andrews pour les détails.

## 12.9 CONCLUSION

L'analyse asymptotique est immanquablement une approximation, puisqu'elle ignore tout ce qui n'est pas de l'ordre dominant par rapport à la taille de l'échantillon. L'analyse de la puissance basée sur la dérive de DGP implique une approximation supplémentaire, puisqu'elle suppose que le DGP

est “proche” de l’hypothèse nulle. Ainsi, bien que les résultats établis dans ce chapitre aient les mérites de la simplicité et d’une application étendue, nous ne pouvons pas attendre d’eux qu’ils fournissent de bonnes approximations dans toutes les situations. En particulier, nous ne pourrions pas espérer des performances de qualité si le DGP était très différent de l’hypothèse nulle.<sup>6</sup> Dans ce cas, bien évidemment, on s’attend à ce que de nombreux tests rejettent l’hypothèse nulle. La plupart des économètres recommenceraient alors sur la base d’un modèle moins contraignant correspondant à une des alternatives contre laquelle le modèle originel a été rejeté, et sans doute plus proche du DGP.

L’objectif de ce chapitre n’est pas de fournir une technique infaillible pour le choix d’un modèle correctement spécifié. Une telle technique n’existe pas. Au lieu de cela, nous avons fourni les éléments d’une structure avec laquelle on peut interpréter les résultats des tests d’hypothèses. L’interprétation d’une statistique de test significative en tant que garantie de validité de l’hypothèse alternative est souvent très exagérée. Il suffit de dénombrer les fois où l’observation d’un  $t$  de Student de 10, par exemple, nous conduit à conclure que le paramètre associé est *définitivement* non nul. Comme nous l’avons vu, cette conclusion est souvent non justifiée. Nous pouvons assurément conclure que le modèle où ce paramètre est nul est mal spécifié, et, dans le cas linéaire, nous pouvons suspecter que la variable associée au paramètre en question est fortement corrélée à tout ce qui est vraiment absent du modèle sous sa forme actuelle. Mais un  $t$  de Student significatif en tant que tel ne nous indique jamais *pourquoi* le modèle est mal spécifié lorsque le paramètre est nul. Par ailleurs, comme nous l’avons vu dans la Section 12.8, une statistique de test non significative n’est pertinente que si le test avait une puissance importante contre des hypothèses économiquement intéressantes.

Dans le prochain chapitre, nous aborderons le thème des tests d’hypothèses, mais dans le contexte de l’estimation par maximum de vraisemblance. La théorie du maximum de vraisemblance offre un support au développement des nombreux tests orientés non-régression, c’est-à-dire des tests qui correspondent à des aspects de la spécification autres que la fonction de régression. Les tests d’hétéroscédasticité dont nous avons discuté dans la Section 11.5 sont des exemples de tels tests; ils sont orientés **fonction scédastique** au lieu d’être orientés régression (voir la Section 16.5). La plupart des résultats restent valables, moyennant une légère modification, pour les tests orientés non-régression autant que pour les tests orientés régression; nous détaillerons tout ceci dans le prochain chapitre. Ils sont également valables pour des modèles estimés à l’aide de procédures GLS et/ou IV.

<sup>6</sup> Nelson et Savin (1990) analysent un exemple simple pour lequel la puissance locale asymptotique d’une statistique de test fournit un indice très mauvais de sa vraie puissance lorsque le DGP diffère quelque peu de l’hypothèse nulle.



## TERMES ET CONCEPTS

convergence (d'un test)	fonction puissance inverse
courbe de niveau-puissance	hypothèse alternative explicite
dérive de DGP	hypothèse alternative implicite
directions de non-régression	hypothèse nulle implicite
directions scédastiques	hypothèse nulle simple
distribution asymptotique (d'une statistique de test)	paramètre de non centralité (NCP)
distribution du $\chi^2$ non centrée	puissance utile
efficacité asymptotique relative (ARE)	suites d'alternatives locales
fonction puissance	test biaisé
	tests asymptotiquement équivalents

# Chapter 13

## Les Tests d'Hypothèses Classiques

### 13.1 INTRODUCTION

Nous avons rencontré pour la première fois au Chapitre 3 les tests d'hypothèses basés sur les principes de LM, Wald, et LR. Cependant, les **trois statistiques de test classiques** elles-mêmes, souvent rattachées irrévérencieusement à la “Sainte Trinité”, n’ont pas été introduits jusqu’à la Section 8.9 parce que, si les tests doivent être appelés *classiques*, ils doivent être exécutés dans le contexte de l’estimation (ML) par maximum de vraisemblance. Comme nous le soulignons au Chapitre 8, l’estimation ML nous impose un dispositif plus restrictif que pour l’estimation NLS ou IV, parce que les DGP du modèle estimés doivent être complètement caractérisés par les paramètres du modèle. Ceci implique que nous devons poser de fortes hypothèses de distribution si nous désirons utiliser l’estimation ML. En retour, ML nous permet d’estimer une variété beaucoup plus large de modèles que ne le permet NLS. De plus, comme nous le verrons dans ce Chapitre, les *tests* basés sur les estimations ML sont beaucoup plus largement applicables que ceux utilisés dans le contexte NLS. Ceci signifie que nous serons capables de construire des tests dans des directions autres que celles des régressions étudiées en détail dans le dernier chapitre.

Heureusement, l’utilisation de ML ne nous oblige pas à abandonner l’utilisation des régressions artificielles. Bien que la régression de Gauss-Newton, que nous avons beaucoup utilisée dans le contexte de l’estimation par moindres carrés et IV, n’est généralement pas applicable aux modèles estimés par ML, nous introduisons une autre régression artificielle à la Section 13.7 qui est la suivante. Il s’agit de la **régression du produit extérieur au gradient**, ou **régression OPG**. La régression OPG peut être utilisée pour l’estimation de la matrice de covariance, test d’hypothèses, estimation efficace en une étape, et ainsi de suite, dans le contexte ML, exactement de la même manière que la GNR peut être utilisée dans le contexte ML et IV. Plus loin dans le livre, nous rencontrerons d’autres régressions artificielles, qui se conduisent habituellement mieux mais qui sont largement moins applicables que la régression OPG.

Ce chapitre est organisé comme suit. La prochaine section fournit une discussion géométrique des trois statistiques de test classiques. La Section

13.3 démontrera aussi qu'ils sont équivalents asymptotiquement sous certaines conditions. La Section 13.4 traite du cas spécial des modèles de régression linéaire et montre comment les statistiques de test classiques sont reliés aux statistiques familières  $t$  et  $F$ . La Section 13.5 discute des moyens variés à partir desquels la matrice d'information peut être estimée, et comment celle-ci affecte les statistiques LM et Wald qui utilisent ces estimations. La Section 13.6 porte sur un résultat important de reparamétrisation et sur comment celui-ci affecte les statistiques de test classiques. Il introduit également le concept d'hypothèses alternatives localement équivalentes. La Section 13.7 introduit la régression OPG et discute brièvement des tests  $C(\alpha)$ . Enfin, certaines suggestions sont fournies pour une lecture ultérieure à la Section 13.8.

## 13.2 LA GÉOMÉTRIE DES STATISTIQUES DE TEST CLASSIQUES

Nous commençons, comme à la Section 8.9, avec un modèle de reparamétrisation que nous appelons le **modèle non contraint** et considérons les restrictions sur ses paramètres, qui définissent implicitement le **modèle contraint**. L'hypothèse nulle alternative ou maintenue correspond au modèle non contraint qui est vrai. Ces deux modèles sont caractérisés par une fonction de logvraisemblance de la forme (8.10), qui est, une somme des contributions des observations dans l'échantillon. Ainsi, pour le modèle non contraint, nous avons une fonction de logvraisemblance pour un échantillon de taille  $n$ :

$$\ell(\mathbf{y}^n, \boldsymbol{\theta}) = \sum_{t=1}^n \ell_t(\mathbf{y}^t, \boldsymbol{\theta}). \quad (13.01)$$

Souvenons-nous que  $\mathbf{y}^t$  désigne un **échantillon de taille  $t$** , qui est, un vecteur d'éléments  $y_s$ ,  $s = 1, \dots, t$ ; ces éléments des vecteurs plutôt que des scalaires, mais nous les traiterons comme des scalaires pour notre notation. Notons que  $\ell_t$  dépend de  $\mathbf{y}^t$  plutôt que de  $y_t$  simplement, parce qu'il peut exister des variables dépendantes retardées ou d'autres formes de dépendances parmi les  $y_t$ . Nous supposons sans plus de commentaire que n'importe quel modèle de la forme (13.01) satisfait les conditions de régularité fournies au Chapitre 8 pour assurer l'existence, la convergence, la normalité asymptotique, et l'efficacité asymptotique de l'estimateur ML pour le modèle. L'invariance de ces estimateurs sous des reparamétrisations du modèle implique, que nous pouvons, utiliser des reparamétrisations arrangeantes afin d'obtenir certains résultats.

La série des DGP générée comme le vecteur paramétrique  $\boldsymbol{\theta}$  varie sur un espace paramétrique  $\Theta_1 \subseteq \mathbb{R}^k$  que constitue le modèle non contraint, que nous désignerons par  $\mathbb{M}_1$ . L'hypothèse alternative est satisfaite par une certaine série de données si les données sont en fait générées par un DGP appartenant à  $\mathbb{M}_1$ . Le modèle contraint, ou hypothèse nulle,  $\mathbb{M}_0$ , représente une sous-série du modèle non contraint  $\mathbb{M}_1$ . Il est généré à partir du (13.01) par l'imposition de restrictions de la forme

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}, \quad \text{où } \mathbf{r} : \Theta_1 \rightarrow \mathbb{R}^r, \quad r < k. \quad (13.02)$$

Nous supposons que les fonctions  $\mathbf{r}(\boldsymbol{\theta})$  qui expriment les restrictions sont **arrangeantes** en  $\boldsymbol{\theta}$  et aussi que l'espace paramétrique  $\Theta_0$  associé avec  $\mathbb{M}_0$  est un sous-espace  $\Theta_1$  arrangeant de dimension  $(k - r)$ . L'hypothèse nulle est satisfaite par une série particulière de données si les données sont générées par un vecteur paramétrique dans le sous-espace  $\Theta_0$ . Comme au Chapitre 8, nous désignerons les **estimations contraintes** par  $\hat{\boldsymbol{\theta}}$  et les **estimations non contraintes** par  $\tilde{\boldsymbol{\theta}}$ .

Le premier test classique est le test LM, qui est basé exclusivement sur les estimations contraintes  $\hat{\boldsymbol{\theta}}$ . Comme nous l'avons vu à la Section 8.9, il peut être réalisé soit sur les multiplicateurs de Lagrange d'un problème de maximisation contraint, soit sur le **vecteur gradient** (ou **vecteur score**) de la fonction de logvraisemblance. Dans sa forme score, il était donné dans l'équation (8.77):

$$LM \equiv n^{-1} \tilde{\mathbf{g}}^\top \tilde{\mathbf{J}}^{-1} \tilde{\mathbf{g}}. \quad (13.03)$$

Rappelons de (8.13) que le vecteur gradient  $\mathbf{g}(\boldsymbol{\theta})$  est défini comme le vecteur colonne des dérivées partielles de la fonction de logvraisemblance à  $\boldsymbol{\theta}$ , que la matrice d'information  $\mathbf{J}(\boldsymbol{\theta})$  est définie dans (8.20), (8.21), et (8.22), et que  $\tilde{\mathbf{g}}$  et  $\tilde{\mathbf{J}}$  désignent ces quantités évaluées en  $\tilde{\boldsymbol{\theta}}$ . Pour une notation aisée dans ce qui suit, nous écrirons  $\mathbf{I}$  à la place de  $n\mathbf{J}$ , car cela nous évitera d'écrire une grande quantité de certaines puissances explicites de  $n$ . Dans la prochaine section, cependant, certaines des puissances de  $n$  auront besoin d'être restaurées lorsque nous nous embarquerons dans une certaine théorie asymptotique. En utilisant la notation  $\mathbf{I}$ , (13.03) peut être réécrite comme

$$LM \equiv \tilde{\mathbf{g}}^\top \tilde{\mathbf{I}}^{-1} \tilde{\mathbf{g}}. \quad (13.04)$$

Le second test classique, le test de Wald, est basé exclusivement sur les estimations non contraintes  $\tilde{\boldsymbol{\theta}}$ . Comme l'hypothèse nulle nécessite que  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ , cela peut être testé en voyant si  $\mathbf{r}(\tilde{\boldsymbol{\theta}})$  est ou n'est pas significativement différent de zéro. Nous avons vu à l'équation (8.78) que qu'une statistique de test convenable est

$$W = n \hat{\mathbf{r}}^\top (\hat{\mathbf{R}} \hat{\mathbf{J}}^{-1} \hat{\mathbf{R}}^\top)^{-1} \hat{\mathbf{r}} = \hat{\mathbf{r}}^\top (\hat{\mathbf{R}} \hat{\mathbf{I}}^{-1} \hat{\mathbf{R}}^\top)^{-1} \hat{\mathbf{r}}, \quad (13.05)$$

où  $\mathbf{R}(\boldsymbol{\theta}) \equiv D_{\boldsymbol{\theta}} \mathbf{r}(\boldsymbol{\theta})$ , et les chapeaux sur  $\mathbf{r}$  et  $\mathbf{R}$  signifient, comme d'habitude, que ces quantités doivent être évaluées en  $\hat{\boldsymbol{\theta}}$ .

La troisième et dernière statistique de test classique est la statistique du ratio de vraisemblance. Elle a été définie dans (8.68):

$$LR = 2(\hat{\ell} - \tilde{\ell}), \quad (13.06)$$

où encore  $\hat{\ell} \equiv \ell(\hat{\boldsymbol{\theta}})$  et  $\tilde{\ell} \equiv \ell(\tilde{\boldsymbol{\theta}})$ .

Afin d'examiner les relations parmi les statistiques de test classiques, nous réalisons tout d'abord une approximation simplificatrice. Il s'agit que le

Hessien de la fonction de vraisemblance soit *constant* dans le voisinage entier de son maximum. Cette approximation est équivalente à supposer que la fonction de logvraisemblance est une fonction quadratique. Si l'approximation était exactement vraie, alors la fonction de logvraisemblance pourrait être écrite comme

$$\ell(\boldsymbol{\theta}) = \hat{\ell} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top n\mathcal{H}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

où la matrice  $\mathcal{H}$ , que désigne le Hessien divisé par  $n$ , est constante, définie positive, indépendante de  $\boldsymbol{\theta}$ , et  $O(1)$ . Comme le Hessien est constant, il doit être égale au minimum de la matrice d'information pour toutes les tailles d'échantillon et pas just asymptotiquement. Ainsi, nous devons remplacer  $n\mathcal{H}$  par  $-\mathbf{I}$ :

$$\ell(\boldsymbol{\theta}) = \hat{\ell} - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathbf{I}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \quad (13.07)$$

En évaluant cette expression en  $\tilde{\boldsymbol{\theta}}$  et en la subdivisant dans la définition de la statistique LR, (13.06), nous voyons que, lorsque la fonction de logvraisemblance est quadratique, la statistique LR peut être réécrite comme

$$LR = (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \mathbf{I}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}). \quad (13.08)$$

Considérons maintenant la statistique LM. De (13.07), il est facile de voir que le gradient  $\tilde{\mathbf{g}}$  est juste  $-\mathbf{I}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ . Alors, de (13.04), il s'en suit que  $LM$  est égale à  $(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \mathbf{I}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ , qui est simplement l'expression (13.08). Ainsi, nous voyons que les tsatistiques LM et LR sont numériquement égales quand la fonction de logvraisemblance est quadratique.

Prouver que ces deux statistiques sont égales à la statistique de Wald dans ce cas spécial est un tout petit peu plus difficile. Nous commençons par poser une autre hypothèse, une qui, comme nous le verrons plus tard, n'occasionne pas en fait une quelconque perte de généralité. Ainsi, les restrictions associées à l'hypothèse nulle prennent la forme

$$\boldsymbol{\theta}_2 = \mathbf{0}. \quad (13.09)$$

Ici nous avons partitionné le vecteur paramétrique comme  $\boldsymbol{\theta} = [\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2]$ , avec  $\boldsymbol{\theta}_2$  un vecteur  $r$ - et  $\boldsymbol{\theta}_1$  par conséquent un vecteur  $(k-r)$ -. Nous pouvons aussi partitionner la matrice d'information de façon à la conformer à la partition de  $\boldsymbol{\theta}$ :

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix}.$$

Avec  $\boldsymbol{\theta}$  et  $\mathbf{I}$  partitionnés de cette manière, l'expression (13.07) pour la fonction de logvraisemblance devient

$$\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \hat{\ell} - \frac{1}{2} \begin{bmatrix} \boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \\ \boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1 \\ \boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2 \end{bmatrix}. \quad (13.10)$$

Au niveau du MLE contraint,  $(\tilde{\boldsymbol{\theta}}_1, \mathbf{0})$ , la condition du premier ordre pour un maximum contraint doit être satisfaite. En différenciant (13.10) par rapport à  $\boldsymbol{\theta}_1$  et en évaluant le résultat à  $\boldsymbol{\theta}_2 = \mathbf{0}$ , nous trouvons que cette condition du premier ordre est

$$\mathbf{0} = D_1 \ell(\tilde{\boldsymbol{\theta}}_1, \mathbf{0}) = -(\mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) - \mathbf{I}_{12}\hat{\boldsymbol{\theta}}_2).$$

De ceci il s'en suit que

$$\mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) = \mathbf{I}_{12}\hat{\boldsymbol{\theta}}_2. \quad (13.11)$$

Si nous écrivons la statistique LR (13.08) sous la forme partitionnée, nous obtenons

$$\begin{aligned} LR &= (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \mathbf{I} (\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \\ &= \begin{bmatrix} \tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1 \\ \tilde{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1 \\ \tilde{\boldsymbol{\theta}}_2 - \hat{\boldsymbol{\theta}}_2 \end{bmatrix} \\ &= (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^\top \mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) - 2(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^\top \mathbf{I}_{12}\hat{\boldsymbol{\theta}}_2 + \hat{\boldsymbol{\theta}}_2^\top \mathbf{I}_{22}\hat{\boldsymbol{\theta}}_2. \end{aligned}$$

où la dernière ligne utilise le fait que  $\tilde{\boldsymbol{\theta}}_2 = \mathbf{0}$ . En utilisant le résultat (13.11), la statistique LR peut être réécrite comme

$$\begin{aligned} LR &= (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^\top \mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) - 2(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^\top \mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1) + \hat{\boldsymbol{\theta}}_2^\top \mathbf{I}_{22}\hat{\boldsymbol{\theta}}_2 \\ &= \hat{\boldsymbol{\theta}}_2^\top \mathbf{I}_{22}\hat{\boldsymbol{\theta}}_2 - (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^\top \mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1). \end{aligned} \quad (13.12)$$

Nous montrons à présent que la statistique de Wald est égale à (13.12). Comme les restrictions prennent la forme (13.09), nous voyons que  $\mathbf{r}(\boldsymbol{\theta}) = \boldsymbol{\theta}_2$  and  $\hat{\mathbf{r}} = \hat{\boldsymbol{\theta}}_2$ . Ceci implique que la matrice  $\mathbf{R}$  peut être écrite comme

$$\mathbf{R}(\boldsymbol{\theta}) = [\mathbf{0} \quad \mathbf{I}],$$

où la matrice  $\mathbf{0}$  est  $r \times (k - r)$ , et la matrice identité  $\mathbf{I}$  est  $r \times r$ . Alors l'expression  $\hat{\mathbf{R}}\mathbf{I}^{-1}\hat{\mathbf{R}}^\top$  qui apparaît dans la statistique de Wald (13.05) est juste le bloc  $(2, 2)$  de la matrice inverse partitionnée  $\mathbf{I}^{-1}$ . Au moyen des résultats de l'Annexe A sur les matrices partitionnées, nous obtenons

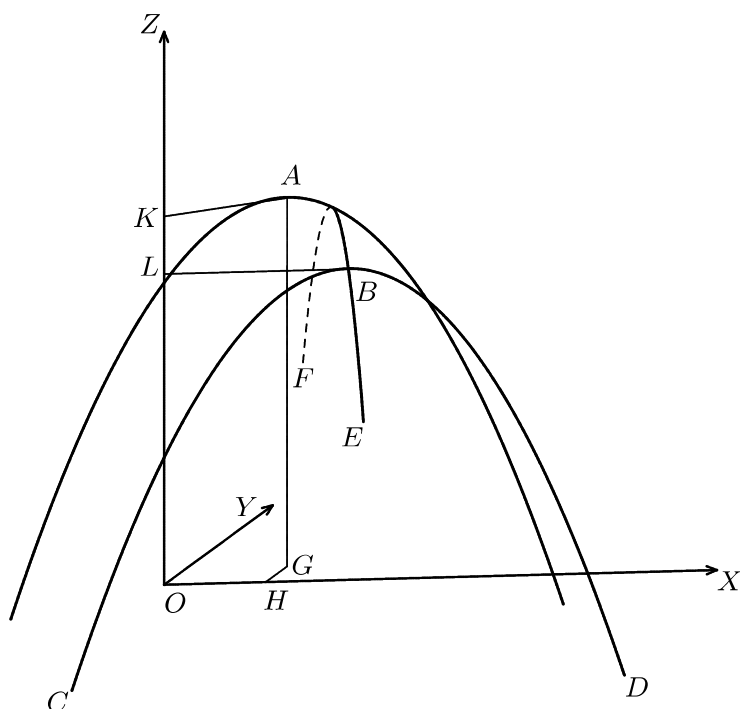
$$(\hat{\mathbf{R}}\mathbf{I}^{-1}\hat{\mathbf{R}}^\top)^{-1} = ((\mathbf{I}^{-1})_{22})^{-1} = \mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12}. \quad (13.13)$$

Ce résultat nous permet de mettre (13.05) sous la forme

$$W = \hat{\boldsymbol{\theta}}_2^\top (\mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12})\hat{\boldsymbol{\theta}}_2.$$

Par (13.11), cette dernière expression est égale à

$$\hat{\boldsymbol{\theta}}_2^\top \mathbf{I}_{22}\hat{\boldsymbol{\theta}}_2 - (\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1)^\top \mathbf{I}_{11}(\tilde{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_1),$$



**Figure 13.1** Maximisation de la fonction de vraisemblance

qui est la même que (13.12). La preuve de l'égalité des trois statistiques classiques pour la fonction de logvraisemblance quadratique (13.07) est par conséquent complète.

Il est intéressant de voir comment les trois statistiques de test classiques sont reliées géométriquement. La Figure 13.1 dépeint le graphe d'une fonction logvraisemblance  $\ell(\mathbf{y}, \theta_1, \theta_2)$ . Ceci est dessiné pour un vecteur échantillon donné  $\mathbf{y}$  et par conséquent pour une taille d'échantillon donnée  $n$ . Pour faire simple, l'espace paramétrique est supposé être à deux dimensions. Il existe seulement une restriction, qui est que le second élément du vecteur paramétrique,  $\theta_2$ , est égal à zéro. Ainsi, la fonction  $\ell$  peut être traitée comme une fonction de deux variables  $\theta_1$  et  $\theta_2$  seulement, et son graphe représenté dans un espace tridimensionnel. Comme d'habitude, la Figure 13.1 est seulement une projection bi-dimensionnelle de cette représentation. Le plan  $(\theta_1, \theta_2)$  devrait être visualisé à l'horizontal, et l'axe vertical,  $OZ$ , mesure ainsi les valeurs de la fonction  $\ell$ .

La fonction de vraisemblance  $\ell$  réalise son maximum  $\hat{\ell}$  au point  $A$  de la figure. Nous pouvons dire que  $A$  possède les coordonnées  $(\hat{\theta}_1, \hat{\theta}_2, \hat{\ell})$  dans le système de coordonnées défini par les trois axes  $OX$ ,  $OY$ , et  $OZ$ . En général,  $\theta_2$  ne sera pas zéro. Si la restriction que  $\theta_2$  est zéro est imposée, alors à la place du maximum  $\ell$  sur le plan entier  $(\theta_1, \theta_2)$ , nous sommes contraints à l'axe  $\theta_1$  et par conséquent à la courbe marquée  $CBD$ . Le maximum contraint,  $\tilde{\ell}$ , est

réalisé au point  $B$ , auquel  $\theta_1 = \tilde{\theta}_1$  et, naturellement,  $\theta_2 = 0$ . Les coordonnées de  $B$  sont alors  $(\tilde{\theta}_1, 0, \tilde{\ell})$ .

Essayons maintenant de trouver des équivalents géométriques dans la Figure 13.1 pour les quantités qui apparaissent dans les trois statistiques de test classiques. Premièrement, pour  $LM$ , notons que  $\tilde{\mathbf{g}}$  est le vecteur gradient de  $\ell$  en  $B$ , représenté géométriquement par l'inclinaison de la tangente en  $B$  à la courbe  $EBF$ , qui est, la courbe dans le graphe de  $\ell$  qui augmente le plus à pic loin de  $B$ . Pour  $W$ , comme la restriction peut être écrite simplement comme  $\theta_2 = 0$ , nous pouvons mettre  $\mathbf{r} = \theta_2$ , et aussi  $\hat{\mathbf{r}} = \hat{\theta}_2$ . Géométriquement,  $\hat{\theta}_2$  est juste une des coordonnées du maximum global de  $\ell$  en  $A$ , et un manière de le définir (parmi d'autres possibilités) c'est par la longueur du segment horizontal  $GH$ .  $G$  est le point  $(\hat{\theta}_1, \hat{\theta}_2, 0)$ , directement au-dessous du point  $A$ , et  $H$  est la projection de  $G$  sur l'axe  $\theta_1$ , c'est à dire, le point  $(\hat{\theta}_1, 0, 0)$ . En dernier lieu, pour  $LR$ , comme  $\hat{\ell}$  et  $\tilde{\ell}$  sont des coordonnées des points  $A$  et  $B$ , respectivement, la différence  $\hat{\ell} - \tilde{\ell}$  est représentée simplement par la longueur du segment vertical  $KL$  sur l'axe  $OZ$ .

L'équivalence des trois statistiques de test classiques peut maintenant être comprise en termes de la géométrie des sommets des collines. Retenons pour le moment l'hypothèse que la fonction de logvraisemblance est exactement quadratique dans le voisinage de son maximum. Afin de simplifier l'algèbre dont nous avons besoin pour exprimer la géométrie, nous posons le changement de variables suivant dans l'espace paramétrique:<sup>1</sup>

$$\begin{aligned}\psi_1 &= \mathbf{I}_{11}^{1/2}(\theta_1 - \hat{\theta}_1) + \mathbf{I}_{11}^{-1/2}\mathbf{I}_{12}(\theta_2 - \hat{\theta}_2); \\ \psi_2 &= (\mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12})^{1/2}(\theta_2 - \hat{\theta}_2).\end{aligned}\tag{13.14}$$

La forme particulière de ce changements de variables est motivée par le fait que la fonction de logvraisemblance, lorsqu'elle est exprimée en termes des  $\psi$ , prend une forme très simple. Premièrement, notons que

$$\psi_1^2 + \psi_2^2 = \begin{bmatrix} \theta_1 - \hat{\theta}_1 \\ \theta_2 - \hat{\theta}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix} \begin{bmatrix} \theta_1 - \hat{\theta}_1 \\ \theta_2 - \hat{\theta}_2 \end{bmatrix},$$

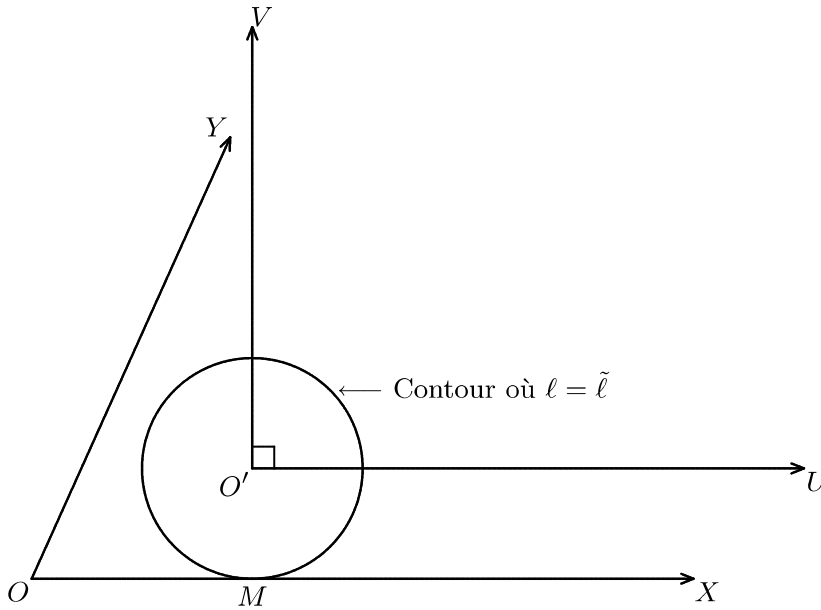
peut être facilement enrayée. Alors il s'en suit de (13.10) que la fonction de logvraisemblance en termes des  $\psi$  est

$$\ell(\psi_1, \psi_2) = \hat{\ell} - \frac{1}{2}(\psi_1^2 + \psi_2^2).\tag{13.15}$$

Par un léger abus de notation, nous continuons à écrire  $\ell$  pour la fonction de logvraisemblance en termes des nouvelles variables.

<sup>1</sup> Nous ne pouvons pas ici parler de reparamétrisation, car le changement de variables est *aléatoire* en raison de sa dépendance sur les estimations paramétriques contraintes





**Figure 13.2** Les systèmes de coordonnées  $\theta$  et  $\psi$

Evidemment, l'effet du changement de variables a été localisé au maximum non contraint de la fonction de logvraisemblance à l'origine des coordonnées  $\psi$  et pour rendre le sommet parfaitement symétrique concernant cette origine. Pour trouver les coordonnées  $\psi$  du maximum contraint, nous pouvons substituer  $\hat{\theta}_1$  et 0 pour  $\theta_1$  et  $\theta_2$  dans (13.14). Nous trouvons de  $\psi_1$  que

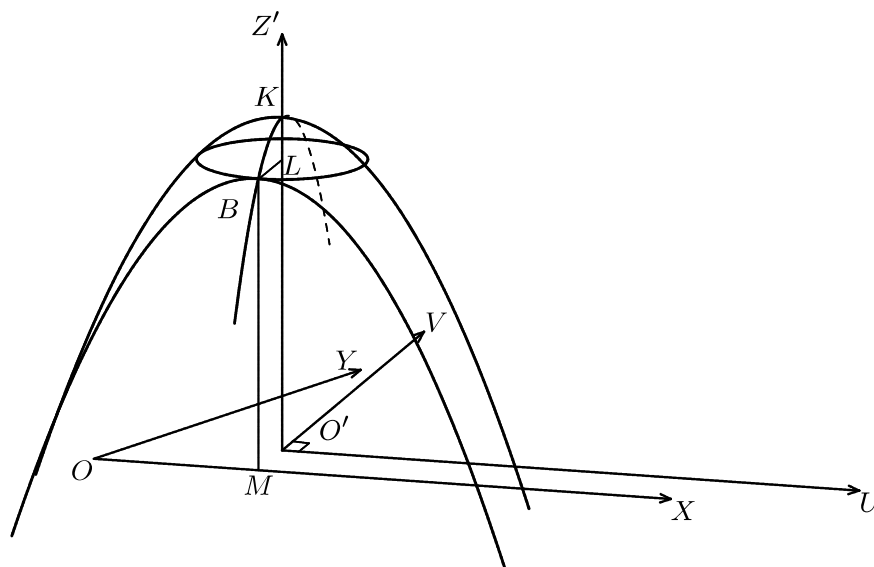
$$\mathbf{I}_{11}^{1/2} \psi_1 = \mathbf{I}_{11}(\tilde{\theta}_1 - \hat{\theta}_1) - \mathbf{I}_{12}\hat{\theta}_2 = 0, \quad (13.16)$$

par (13.11), qui implique que la coordonnée  $\psi_1$  du maximum contraint est zéro. Ce fait peut être exprimé de façon plus géométrique en disant que le maximum contraint est atteint à un point sur l'axe  $\psi_2$ . Pour la coordonnée  $\psi_2$ , le résultat est

$$\psi_2 = -(\mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12})^{1/2} \hat{\theta}_2. \quad (13.17)$$

La restriction  $\theta_2 = 0$  est satisfaite en un point dans l'espace paramétrique si et seulement si (13.17) est satisfait par la coordonnée  $\psi_2$  du point. Ceci signifie que, en termes des  $\psi$ , non seulement le maximum contraint mais aussi la série entière des vecteurs paramétriques correspond aux DGP qui satisfont l'hypothèse nulle qui s'établit sur la ligne directe, parallèle à l'axe  $\psi_1$ , avec l'équation (13.17).

Comme nous l'avons remarqué plus tôt, (13.15) montre que le sommet de la colline réalisé par la fonction de logvraisemblance est parfaitement symétrique concernant l'origine dans les coordonnées  $\psi$ . Redessinons la Figure 13.1 en utilisant  $\psi$  à la place de  $\theta$ . Dans la Figure 13.2, seul l'espace paramétrique a été dessiné. Deux séries des axes ont été superposés. Les axes  $\psi$  sont dessinés de façon orthogonale les uns aux autres de manière habituelle,



**Figure 13.3** La logvraisemblance en fonction de  $\psi_1$  et  $\psi_2$

mais ce fait implique que les axes  $\theta$  ne peuvent pas en général être mutuellement orthogonaux. (Les axes  $\psi$  reçoivent ce traitement privilégié parce qu'ils rendent seulement la fonction de logvraisemblance symétrique par rapport à l'origine.) Ensuite, la Figure 13.3 montre l'image complète en trois dimensions. La nouvelle origine,  $O'$ , est localisée à l'ancien  $(\hat{\theta}_1, \hat{\theta}_2)$ , au-dessous du maximum de  $\ell$ .  $\psi_1$  axis, dessiné comme la ligne  $O'U$ , est parallèle à l'ancien axe  $\theta_1$ ,  $OX$ . Ceci provient du fait que l'axe  $\theta_1$  est la série des vecteurs paramétriques satisfaisant l'hypothèse nulle et de notre observation précédente que cette série coïncide avec la ligne (13.17) parallèle à l'axe  $\psi_1$ ? revoir cette dernière phrase? . L'axe  $\psi_2$ ,  $O'V$ , est perpendiculaire à  $O'U$  mais pas parallèle en général à l'axe  $\theta_2$   $OY$ .

Une conséquence de la forme symétrique de (13.15) est que les courbes de niveau de la fonction  $\ell$  sont devenues des *cercles* centrés sur l'origine  $\psi$  dans les nouvelles figures. Nous avons vu de (13.16) que le maximum contraint de  $\ell$  est réalisé sur l'axe  $\theta_1$   $OX$  ou pour pour lequel  $\psi_1 = 0$ , qui est, au point  $M$  où il croise l'axe  $\psi_2$ . Par un raisonnement standard, nous pouvons voir que la courbe de niveau de  $\ell$  sur laquelle repose le maximum contraint, pour laquelle  $\ell = \tilde{\ell}$ , doit être *tangente* à l'axe  $\theta_1$  et ainsi aussi à l'axe  $\psi_1$  au maximum; voir la Figure 13.2. Le rayon  $O'M$  de la courbe de niveau circulaire  $\ell = \tilde{\ell}$  qui joins l'origine  $\psi$  au point de tangence est perpendiculaire à la tangente, qui explique géométriquement pourquoi le maximum contraint repose sur l'axe  $\psi_2$ . La longueur du rayon  $O'M$  est, naturellement, juste la valeur de  $\psi_2$  donné par (13.17), que nous désignerons par  $\rho$ .

Rappelons maintenant que la statistique LR était représentée géométriquement par le segment de ligne vertical  $KL$  dans les Figures 13.1 et 13.3. Nous

pouvons utiliser (13.06) et (13.15) directement pour obtenir la longueur de ce segment :

$$LR = 2(\hat{\ell} - \ell(0, \rho)) = \rho^2.$$

Pour la statistique LM, il est clair que le gradient de  $\ell$ , par rapport aux coordonnées  $\psi$ , en n'importe quel point  $(\psi_1, \psi_2)$  est le vecteur  $-(\psi_1, \psi_2)$ . En  $M$  il est par conséquent juste  $-(0, \rho)$ . Plus loin, le Hessien de  $\ell$  par rapport aux coordonnées  $\psi$  en n'importe quel point  $(\psi_1, \psi_2)$  est simplement moins la matrice identité  $2 \times 2$ . Ainsi, si nous utilisons le gradient par rapport à  $\psi$  dans (13.04), le long de la négative du Hessien par rapport au  $\psi$  en place dans la matrice d'information, nous obtenons une statistique LM égale au <sup>2</sup>

$$\begin{bmatrix} 0 & \rho \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \rho \end{bmatrix} = \rho^2.$$

Rappelons maintenant que le test de Wald correspond géométriquement au segment de ligne  $GH$  dans la Figure 13.1, qui est devenue le rayon  $O'M$  dans la Figure 13.3, de longueur  $\rho$ . Avant le changement de variables des  $\theta$  en  $\psi$ , la statistique de Wald devait avoir été calculée comme le carré de la longueur de  $GH$  divisé par une estimation appropriée de la variance de  $\hat{\theta}_2$ . Comme dans la Figure 13.3 les axes  $\psi$  sont mutuellement orthogonaux plutôt que les axes  $\theta$ , la longueur de  $O'M$  dans cette figure est différente de la longueur de  $GH$  dans la Figure 13.1. Ainsi nous devrions utiliser une mesure de variance différente lorsque nous travaillons en termes des  $\psi$ . Juste comme nous l'avons fait pour la statistique LM, nous pouvons utiliser la matrice identité  $2 \times 2$  à la place de la matrice d'information. Ceci signifie que la mesure de variance appropriée est juste l'unité, et comme la longueur de  $O'M$  est  $\rho$ , la statistique de Wald est juste  $\rho^2$ .

Notre preuve précédente de l'égalité des trois statistiques de test classiques fournit une justification des arguments heuristiques d'avant. Cependant, nous pouvons vérifier directement que la quantité de  $\rho^2$  est que nous avons obtenu en travaillant avec les  $\theta$  ? revoir la phrase?. Il s'en suit directement de (13.17) que

$$\rho^2 = (\mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12})\hat{\theta}_2^2.$$

Par (13.11) appliqué au cas présent à deux paramètres, ceci devient

$$\rho^2 = \mathbf{I}_{22}\hat{\theta}_2^2 - \mathbf{I}_{11}(\tilde{\theta}_1 - \hat{\theta}_1)^2.$$

C'est que l'expression (13.12) pour les réductions des trois statistiques de test classiques quant  $\theta_1$  et  $\theta_2$  sont scalaires.

<sup>2</sup> Cet argument est heuristique, car, à strictement parlé, nous ne serions pas en train de traiter les  $\psi$  comme s'ils constituaient des paramètres ordinaires. Cependant, nous procéderons comme s'il était possible de faire ainsi.

Ce que l'utilisation des  $\psi$  nous permet de voir clairement, en termes du sommet symétrique généré par (13.15), est juste pourquoi les trois tests sont équivalents dans le cas présent simple. Tous les trois mesurent, en un certain sens, de combien le MLE non contraint est loin du MLE contraint. Le test de Wald est basé directement sur la distance entre ces deux points dans l'espace paramétrique. Géométriquement, c'est la longueur de  $\rho$  du rayon  $O'M$  de la courbe de niveau circulaire de la fonction de logvraisemblance pour la valeur  $\tilde{\ell}$ . La distance entre les deux estimations, pour les objectifs du test de Wald, est par conséquent, \* mesurée par la distance Euclidienne au carré en termes de  $\psi$  entre les vecteurs paramétriques auxquels les maxima contraints et non contraints sont réalisés. Cela ne serait pas vrai avec les  $\theta$ , naturellement, car cela demande que le sommet soit symétrique. Pour le test du ratio de vraisemblance, la mesure de distance est en termes de la réelle différence entre les deux maxima. Ainsi, géométriquement, la statistique LR est reliée à la longueur d'un segment vertical,  $KL$  dans les figures, tandis que la statistique est reliée à la longueur du segment de ligne horizontal. Pour finir, la statistique LM est basée sur l'inclinaison du chemin le plus pentu de la colline en l'estimation contrainte.

Pour un sommet parfaitement symétrique, ce que nous avons montré dans cette section est que toutes les mesures des trois distances sont fonction de la longueur  $\rho$  du rayon de la courbe de niveau passant à travers le maximum contraint seulement et aussi sont exactement équivalentes. Ce que nous verrons plus tard dans ce chapitre est que ce résultat agréable est exactement vrai *seulement* lorsque le sommet de la logvraisemblance est exactement quadratique, c'est, lorsque le Hessien de la fonction de logvraisemblance est en effet exactement constante dans le voisinage entier de son maximum. Mais tous les collines sont *rudement* quadratiques en leurs sommets, et dans la prochaine section nous serons capables d'exploiter ce fait pour démontrer que toutes ces trois statistiques de test classiques sont *asymptotiquement* équivalentes sous des conditions de régularité étroites.

### 13.3 L'EQUIVALENCE ASYMPTOTIQUE DES TESTS CLASSIQUES

Dans cette section, nous établissons deux séries de résultats concernant l'équivalence asymptotique des trois tests classiques. La première, et plus faible, série est dérivée sous l'hypothèse que les restrictions sur les tests sont en fait vrai. Plus formellement, nous supposons que les données sont générées par un DGP caractérisé par un vecteur paramétrique  $\theta_0$  qui satisfait  $r(\theta_0) = \mathbf{0}$ , dans la notation de (13.02). Notons que nous supposons peu de temps que les restrictions prennent la forme spéciale (13.09). L'équivalence des trois statistiques de test classiques dans ce cas est maintenant absolument simple à démontrer, car les ingrédients principaux ont déjà été établis au Chapitre 8. Notre travail consiste ici principalement à assembler les pièces ensemble et à vérifier que l'intuition du cas exactement quadratique traité dans cette

dernière section s'étend en effet à un résultat asymptotiquement valide en général.

La seconde, et plus forte, série de résultats ne sera pas établie en détail dans ce livre. Ces résultats s'étendent à ceux de la première série au cas dans lequel les données sont générées par un DGP en mouvement qui ne satisfait pas l'hypothèse nulle mais tend vers une limite contenue en lui. Ainsi, cette seconde série de résultats est analogue au cas de l'estimation ML pour les résultats obtenus au Chapitre 12 pour l'estimation par NLS. Bien que nous ne fournirons pas de preuves entières, nous prendrons un peu de temps pour établir les résultats et expliquer ce que nous voulons dire par DGP en mouvement dans ce nouveau contexte.

Pour la première série de résultats, alors, nous supposons que le véritable vecteur paramétrique  $\theta_0$  obéit à (13.02). dans ce cas, à la fois le MLE non contraint  $\hat{\theta}$  et le MLE contraint  $\tilde{\theta}$  diffèrent de  $\theta_0$  par une quantité aléatoire de l'ordre de  $n^{-1/2}$ . Pour le MLE non contraint,

$$n^{1/2}(\hat{\theta} - \theta_0) \stackrel{a}{=} \mathcal{J}_0^{-1} n^{-1/2} g_0, \quad (13.18)$$

un résultat qui suit immédiatement de (8.38) et de l'égalité de la matrice d'information que  $\mathcal{J}_0 = -\mathcal{H}_0$ . Comme nous sommes encore dans le contexte de la théorie asymptotique, nous utilisons maintenant la notation avec les puissances explicites de la taille d'échantillon  $n$ . Sinon, la notation habituelle est:  $\mathcal{J}_0$  et  $g_0$  désignent  $\mathcal{J}(\theta_0)$  et  $g(\theta_0)$ , respectivement. Pour les MLE contraints, nous utilisons l'égalité de la matrice d'information et un résultat suivant immédiatement (8.74) pour obtenir

$$n^{1/2}(\tilde{\theta} - \theta_0) \stackrel{a}{=} \mathcal{J}_0^{-1} (\mathbf{I} - \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{J}_0^{-1}) n^{-1/2} g_0. \quad (13.19)$$

La matrice d'information  $\mathcal{J}(\theta)$  est de l'ordre de l'unité et lisse en  $\theta$ , suivant nos hypothèses standards pour les modèles qui doivent être estimés par maximum de vraisemblance, hypothèses qui ont été déployées dans les énoncés des Théorèmes 8.1, 8.2, et 8.3. Il s'en suit que

$$\ddot{\mathcal{J}} = \mathcal{J}_0 + O(n^{-1/2}), \quad (13.20)$$

où  $\ddot{\mathcal{J}} \equiv \mathcal{J}(\ddot{\theta})$ . Ceci demeure vrai pour n'importe quel estimateur convergent au taux  $n^{1/2}$   $\ddot{\theta}$ , mais nous adopterons la convention que  $\ddot{\theta}$  désigne, soit  $\hat{\theta}$ , soit  $\tilde{\theta}$ . De façon similaire, comme les fonctions  $r(\theta)$  qui définissent les restrictions sont supposées être lisses, nous avons

$$\ddot{\mathbf{R}} = \mathbf{R}_0 + O(n^{-1/2}), \quad (13.21)$$

où la relation d'ordre doit être comprise élément par élément dans les deux équations précédentes.

Considérons premièrement la statistique de test LM (13.03). Elle peut être écrite comme

$$LM = (n^{-1/2}\tilde{\mathbf{g}})^\top \tilde{\mathbf{J}}^{-1} (n^{-1/2}\tilde{\mathbf{g}}), \quad (13.22)$$

où les parenthèses mettent en relief le fait qu'il s'agit d'une forme quadratique dans le vecteur  $O(1) n^{-1/2}\tilde{\mathbf{g}}$ . Nous pouvons développer ce vecteur par le Théorème de Taylor pour obtenir

$$n^{-1/2}\tilde{\mathbf{g}} = n^{-1/2}\mathbf{g}_0 + (n^{-1}\mathbf{H}(\bar{\boldsymbol{\theta}}))n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0);$$

consulter (8.35). La loi des grands nombres appliquée à la matrice Hessienne  $\mathbf{H}$  et à l'égalité de la matrice d'information nous permet d'écrire ceci comme

$$n^{-1/2}\tilde{\mathbf{g}} \stackrel{a}{=} n^{-1/2}\mathbf{g}_0 - \mathcal{J}_0 n^{1/2}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Nous sommes en mesure à présent d'employer le résultat (13.19) pour trouver que

$$n^{-1/2}\tilde{\mathbf{g}} \stackrel{a}{=} \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{J}_0^{-1} n^{-1/2}\mathbf{g}_0.$$

La substitution de ceci dans (13.22) donne alors

$$LM \stackrel{a}{=} (n^{-1/2}\mathbf{g}_0)^\top \mathcal{J}_0^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{J}_0^{-1} (n^{-1/2}\mathbf{g}_0). \quad (13.23)$$

Notons que cette expression se compose seulement de quantités évaluées au véritable vecteur paramétrique  $\boldsymbol{\theta}_0$  et que, de celles-ci, seule  $\mathbf{g}_0$  est stochastique.

Pour la statistique du ratio de vraisemblance (13.06), un autre développement de Taylor est nécessaire. Cette fois, en développant autour de  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$  et en utilisant les équations de vraisemblance  $\hat{\mathbf{g}} = \mathbf{0}$ , nous obtenons

$$LR = 2(\hat{\ell} - \tilde{\ell}) \stackrel{a}{=} n(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})^\top \hat{\mathbf{J}}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}); \quad (13.24)$$

consulter aussi (8.70). En combinant (13.18) et (13.19), nous obtenons

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \stackrel{a}{=} \mathcal{J}_0^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{J}_0^{-1} (n^{-1/2}\mathbf{g}_0).$$

En substituant cette dernière relation dans (13.24) et en remplaçant  $\hat{\mathbf{J}}$  par  $\mathcal{J}_0$ , comme (13.20) nous permet de le faire, nous trouvons que

$$LR \stackrel{a}{=} (n^{-1/2}\mathbf{g}_0)^\top \mathcal{J}_0^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{J}_0^{-1} (n^{-1/2}\mathbf{g}_0). \quad (13.25)$$

Comme ceci est identique à l'expression (13.23), nous avons établi l'équivalence asymptotique de  $LM$  et  $LR$ .

Finalement, nous considérons la statistique de Wald (13.05). Le développement de Taylor des restrictions  $\mathbf{r}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ , l'utilisation de l'hypothèse que  $\mathbf{r}(\boldsymbol{\theta}_0) = \mathbf{0}$ , et la multiplication par  $n^{1/2}$ , nous permettent d'obtenir

$$n^{1/2}\hat{\mathbf{r}} \stackrel{a}{=} \mathbf{R}_0 n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \mathbf{R}_0 \mathcal{J}_0^{-1} n^{-1/2}\mathbf{g}_0,$$

où l'égalité finale provient de (13.18). Ce dernier résultat, le long de (13.20) et (13.21), nous permet de réécrire (13.05) asymptotiquement comme

$$W \stackrel{a}{=} (n^{-1/2} \mathbf{g}_0)^\top \mathcal{I}_0^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{I}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{I}_0^{-1} (n^{-1/2} \mathbf{g}_0). \quad (13.26)$$

L'équivalence asymptotique des trois tests classiques sous les hypothèses nulles des tests, est maintenant établie à travers l'égalité de (13.23), (13.25), et (13.26).

A ce point, il est facile de dériver la distribution asymptotique commune des trois statistiques de test classiques. Rappelons de (8.41) que la distribution asymptotique de  $n^{-1/2} \mathbf{g}_0$  est  $N(\mathbf{0}, \mathcal{I}_0)$ . De ceci, nous calculons que la distribution asymptotique du vecteur  $\mathbf{R}_0 \mathcal{I}_0^{-1} (n^{-1/2} \mathbf{g}_0)$  de dimension  $r$  est  $N(\mathbf{0}, \mathbf{R}_0 \mathcal{I}_0^{-1} \mathbf{R}_0^\top)$ . Les trois statistiques de test sont, comme nous venons juste de le voir, égale asymptotiquement égales à la forme quadratique (13.26) dans le vecteur  $\mathbf{R}_0 \mathcal{I}_0^{-1} (n^{-1/2} \mathbf{g}_0)$  et la matrice  $(\mathbf{R}_0 \mathcal{I}_0^{-1} \mathbf{R}_0^\top)^{-1}$ . Il s'en suit immédiatement que la distribution asymptotique des statistiques de test classiques une chi-deux centrale à  $r$  degrés de liberté.

Notre discussion du premier cas, dans lequel les restrictions sous test sont en fait réelles, est à présent complète. Nous tournons par conséquent notre attention sur le second cas, dans lequel les données sont générées par une mise en marche du DGP qui tend à la limite vers un DGP qui satisfait l'hypothèse nulle. Nous commençons par la discussion du concept de mise en marche des DGP dans le contexte des modèles qui doivent être estimés par la méthode du maximum de vraisemblance.

Dans le contexte des modèles estimés par NLS, nous avons obtenu une mise en marche d'un DGP en ajoutant une quantité proportionnelle au taux  $n^{-1/2}$  à la fonction de régression  $\mathbf{x}(\boldsymbol{\beta}_0)$ ; rapellons (12.06). Ainsi, comme  $n \rightarrow \infty$ , le DGP a évolué à un taux convenable vers un spécifié par le vecteur paramétrique  $\boldsymbol{\beta}_0$ , supposé satisfaire les restrictions des hypothèses nulles. Tout comme les modèles NLS sont définis au moyen de leurs fonctions de régression, les modèles qui doivent être estimés par maximum de vraisemblance sont définis au moyen de leurs fonctions de logvraisemblance, comme dans (13.01). Dans le contexte des modèles ML, il semble par conséquent approprié d'ajouter une quantité appropriée au taux  $n^{-1/2}$  à la contribution pour la fonction de logvraisemblance de chaque observation. Ainsi nous écrivons de l'observation  $t$

$$\ell_t = \ell_t(\mathbf{y}^t, \boldsymbol{\theta}_0) + n^{-1/2} a_t(\mathbf{y}^t). \quad (13.27)$$

Nous pouvons voir de ceci que le log de la densité de la  $t^{\text{ième}}$  observation est pris pour être comme donné par un modèle paramétrique pour un vecteur paramétrique  $\boldsymbol{\theta}_0$  satisfaisant les restrictions de l'hypothèse nulle, plus un terme qui disparaît avec  $n^{-1/2}$  quand  $n \rightarrow \infty$ . Le fait que n'importe quelle fonction de densité est normalisée pour intégrer l'unité signifie que les fonctions  $a_t$  dans

(13.27) doivent être choisie de façon à obéir à la condition de normalisation

$$\int \exp(\ell_t + n^{-1/2}a_t) dy_t = 1.$$

il peut être facilement montré que ceci implique que

$$E_0(a_t(\mathbf{y}^t)) = O(n^{-1/2}), \quad (13.28)$$

où  $E_0$  désigne une espérance calculée en utilisant  $\ell_t(\mathbf{y}^t, \boldsymbol{\theta}_0)$  comme densité de log. Alors, afin de conduire l'ordre asymptotiquement, les variables aléatoires  $a_t$  ont une moyenne nulle.

Le fait que  $\ell_t$  est écrit dans (13.27) comme la somme de deux termes ne contraint pas du tout l'application de l'analyse asymptotique, parce que l'on peut penser (13.27) comme la résultante d'une approximation des séries de Taylor pour une quelconque mise en marche d'un DGP. Un exemple serait séquence des alternatives locales

$$\ell_t(\mathbf{y}^t, \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{\delta}).$$

Par des arguments similaires à ceux de la Section 12.3, nous pouvons montrer qu'une approximation des séries de Taylor comme ceci peut être écrite sous la forme de (13.27).

Nous établissons maintenant sans preuve les résultats qui correspondent aux équations (12.11), (12.12), et (12.13) dans le contexte NLS. Ils sont discutés et prouvés dans Davidson et MacKinnon (1987), une étude que certains lecteurs peuvent, cependant, trouver quelque peu difficile en raison de la nature des mathématiques employées. Ces résultats fournissent des expressions valides asymptotiquement pour les éléments variés des statistiques de test classiques sous la mise en marche d'un DGP spécifié par (13.27). Le premier résultat est que les estimateurs  $\hat{\boldsymbol{\theta}}$  et  $\tilde{\boldsymbol{\theta}}$  sont encore convergents au taux  $n^{1/2}$  pour  $\boldsymbol{\theta}_0$ :

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + O(n^{-1/2}),$$

duquel nous pouvons conclure que  $\ddot{\mathbf{J}}$  et  $\ddot{\mathbf{R}}$  sont convergents pour  $\mathbf{J}_0$  et  $\mathbf{R}_0$ , tout comme ils le sont sous l'hypothèse nulle:

$$\ddot{\mathbf{J}} = \mathbf{J}_0 + O(n^{-1/2}); \quad \text{and} \quad \ddot{\mathbf{R}} = \mathbf{R}_0 + O(n^{-1/2}).$$

Nous pouvons également conclure de la convergence de  $\tilde{\boldsymbol{\theta}}$  que tous les développements de Taylor expansions dans les équations de développement (13.23), (13.25), et (13.26) sont encore valides, tout comme le sont ces équations elles-mêmes.

Comme cela peut être vu des équations (13.23), (13.25), et (13.26), la partie stochastique de toutes les statistiques de test classiques, asymptotiquement, est la quantité  $n^{-1/2}\mathbf{g}_0$ . Son comportement n'est pas le même sous la mise en marche d'un DGP qu'il ne l'est sous l'hypothèse nulle; rappelons-nous



(12.13) pour le cas NLS. Nous trouvons que la distribution asymptotique de  $n^{-1/2}\mathbf{g}_0$  est encore normale mais n'a plus de moyenne nulle. Si nous définissons le vecteur  $\mathbf{c}$   $O(1)$  de dimension  $k$ - par

$$\mathbf{c} \equiv \lim_{n \rightarrow \infty} \text{cov}_0 \left( n^{-1/2} \sum_{t=1}^n a_t(\mathbf{y}^t), n^{-1/2} \mathbf{g}_0 \right), \quad (13.29)$$

où  $\text{cov}_0$  signifie une covariance calculée sous le DGP limite caractérisé par  $\boldsymbol{\theta}_0$ , alors il peut être prouvé que, asymptotiquement,

$$n^{-1/2} \mathbf{g}_0 \sim N(\mathbf{c}, \mathcal{J}_0).$$

Par conséquent, la distribution asymptotique du vecteur de dimension  $r$ -  $\mathbf{R}_0 \mathcal{J}_0^{-1}(n^{-1/2} \mathbf{g}_0)$  est maintenant  $N(\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{c}, \mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)$ , et la distribution asymptotique des statistiques de test classiques est une **chi-deux non centrale** à  $r$  degrés de liberté et à paramètre non central

$$\Lambda = \mathbf{c}^\top \mathcal{J}_0^{-1} \mathbf{R}_0^\top (\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top)^{-1} \mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{c}. \quad (13.30)$$

Jusqu'ici, nos résultats sont très similaires à ceux du dernier chapitre pour la mise en marche des DGP dans le contexte d'un modèle de régression. En fait, les similitudes sont même plus profondes que ce qui a été vu jusqu'ici. Presque toute notre discussion de la géométrie de la puissance de test donnée dans la Section 12.5 peut être remplacée, avec seulement quelques légères modifications, pour le cas présent. Une différence qui en premier lieu pourrait apparaître présenter un obstacle insurmontable mais qui en fait très anodin, est que la mise en marche des DGP comme (13.27) doit être placée dans un espace de dimension infini plutôt que dans l'espace de dimension  $n$ - utilisé auparavant. Ceci est ainsi parce que chaque  $a_t$  est une fonction de l'observation  $y_t$ , et parce que les séries de fonctions sont de dimension de type infini. Mais pour notre objectif ceci signifie simplement que les *possibilités* pour construire les DGP qui tendent vers un DGP qui satisfait l'hypothèse nulle sont infinies. En particulier, tout comme dans le contexte présent nous ne sommes ni contraints aux tests dans des directions de régression, , ni contraints de considérer la mise en route des DGP dans des directions de régression.

La géométrie de ce que nous avons réalisé peut être illustrée en trois dimensions, tout comme dans la Figure 12.3. Supposons pour faire simple que nous travaillons sur une reparamétrisation dans laquelle la matrice d'information  $\mathcal{J}_0$  est une matrice identité. ix.<sup>3</sup> Plus loin, supposons qu'il existe seulement deux paramètres,  $\theta_1$  et  $\theta_2$ , et que la restriction correspondant à l'hypothèse nulle est que  $\theta_2 = 0$ . Ainsi

$$\mathcal{J}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{R}_0 = [0 \quad 1].$$

<sup>3</sup> Par exemple, nous pourrions utiliser pour une reparamétrisation (13.14) avec des quantités aléatoires  $\hat{\theta}_1$  et  $\hat{\theta}_2$  remplacées par les véritables valeurs qui correspondent au DGP limite, et  $\mathbf{I}$  remplacée par  $\mathcal{J}$  évaluée en ces valeurs.

Dans ce cas simple, la matrice de covariance  $\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{R}_0^\top$  se réduit au scalaire 1, le vecteur  $\mathbf{R}_0 \mathcal{J}_0^{-1} \mathbf{c}$  devient simplement  $c_2$ , le second élément du vecteur  $\mathbf{c}$  de dimension 2-, et de (13.30) le paramètre non central  $\lambda$  qui devient  $c_2^2$ .

L'espace à trois dimension que nous construirons est engendré par trois variables aléatoires, interprétées comme des vecteurs dans cet espace. Ces variables aléatoires sont précisément celles qui apparaissent dans la définition (13.29) de  $\mathbf{c}$ . Elles sont

$$\begin{aligned} \mathbf{s}_1 &\equiv n^{-1/2} \sum_{t=1}^n \frac{\partial \ell_t}{\partial \theta_1}(\mathbf{y}^t, \boldsymbol{\theta}_0), \\ \mathbf{s}_2 &\equiv n^{-1/2} \sum_{t=1}^n \frac{\partial \ell_t}{\partial \theta_2}(\mathbf{y}^t, \boldsymbol{\theta}_0), \text{ and} \\ \mathbf{a} &\equiv n^{-1/2} \sum_{t=1}^n a_t(\mathbf{y}^t). \end{aligned}$$

Afin de traiter ces variables aléatoires comme des vecteurs dans un espace Euclidien à trois dimensions qu'elles engendrent, il est suffisant de s'assurer que les opérations algébriques définies sur des espaces Euclidiens peuvent être proprement définies pour ces variables aléatoires. Les opérations d'addition et de multiplication par un scalaire sont définies de manière évidente. La somme d'espace-Euclidien, ou de façon plus concise **somme vectorielle**, de deux variables aléatoires est simplement leur somme ordinaire:

$$\mathbf{s}_1 + \mathbf{s}_2 = n^{-1/2} \sum_{t=1}^n \left( \frac{\partial \ell_t}{\partial \theta_1}(\mathbf{y}^t, \boldsymbol{\theta}_0) + \frac{\partial \ell_t}{\partial \theta_2}(\mathbf{y}^t, \boldsymbol{\theta}_0) \right). \quad (13.31)$$

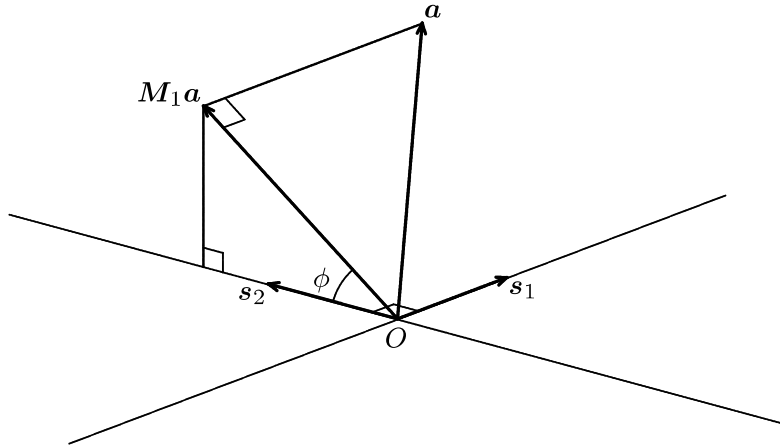
De manière similaire, la multiplication par un scalaire  $\alpha \in \mathbb{R}$  n'est pas différente dans le contexte de l'espace-Euclidien de la multiplication par un scalaire:

$$\alpha \mathbf{s}_1 = \alpha n^{-1/2} \sum_{t=1}^n \frac{\partial \ell_t}{\partial \theta_1}(\mathbf{y}^t, \boldsymbol{\theta}_0). \quad (13.32)$$

Ces deux définitions (13.31) et (13.32) suffisent à placer la structure de l'espace linéaire  $\mathbb{R}^3$  sur la série de toutes les combinaisons linéaires de  $\mathbf{s}_1$ ,  $\mathbf{s}_2$ , et  $\mathbf{a}$ .

Pour un espace Euclidien, une autre opération doit être définie, à savoir, le **produit intérieur** de deux vecteurs. Ainsi, nous espérons être capables de dire que ce que nous signifions par produit intérieur de n'importe quelle combinaison linéaire de  $\mathbf{s}_1$ ,  $\mathbf{s}_2$ , et  $\mathbf{a}$  avec n'importe quelle autre combinaison linéaire. Ceci est réalisé très simplement: le produit intérieur de deux variables aléatoires sera leur covariance sous le DGP limite. Nous désignerons les produits intérieurs par une notation entre crochets d'angles dans laquelle, par exemple,  $\langle \mathbf{s}_1, \mathbf{s}_2 \rangle$  désigne le produit intérieur de  $\mathbf{s}_1$  avec  $\mathbf{s}_2$ . Comme la matrice de covariance de  $\mathbf{s}_1$  et de  $\mathbf{s}_2$ ,  $\mathcal{J}_0$ , sont des matrices identité, nous avons

$$\langle \mathbf{s}_1, \mathbf{s}_2 \rangle = 0 \quad \text{and} \quad \langle \mathbf{s}_i, \mathbf{s}_i \rangle \equiv \|\mathbf{s}_i\|^2 = 1, \quad i = 1, 2.$$



**Figure 13.4** Le paramètre de non centralité des tests classiques

Il s'en suit que  $\mathbf{s}_1$  et  $\mathbf{s}_2$  forment une paire de vecteurs unitaires mutuellement orthogonaux. Notons que la norme au carré d'un élément de notre espace Euclidien de variables aléatoires, définie, comme d'habitude, comme le produit intérieur de l'élément avec lui-même, est juste le second moment de l'élément considéré comme une variable aléatoire. Comme les variables  $\mathbf{s}_1$  et  $\mathbf{s}_2$  ont une espérance nulle sous le DGP limite, la norme au carré autre que la leur est juste la variance. Asymptotiquement, la même chose est vraie pour  $\mathbf{a}$ , car par (13.28) l'espérance de  $\mathbf{a}$  disparaît quand  $n \rightarrow \infty$ .

En général, le vecteur représenté par la variable aléatoire  $\mathbf{a}$  ne tiendra pas dans le plan engendré par  $\mathbf{s}_1$  et  $\mathbf{s}_2$ ; c'est pourquoi une troisième dimension est nécessaire afin d'accomoder cela. Considérons maintenant la Figure 13.4, dans laquelle les vecteurs  $\mathbf{s}_1$  et  $\mathbf{s}_2$  engendrent le plan horizontal. En suivant l'intuition du dernier chapitre, nous laissons  $\mathbf{M}_1\mathbf{a}$  être la projection de  $\mathbf{a}$  sur le complément orthogonal l'espace à une dimension engendré par  $\mathbf{s}_1$ . Il s'en suit que  $\mathbf{M}_1\mathbf{a}$  repose verticalement au-dessus ou au-dessous de la direction de  $\mathbf{s}_2$  (nous l'avons dessiné au-dessus). L'angle désigné  $\phi$  dans la figure est l'angle compris entre le vecteur  $\mathbf{M}_1\mathbf{a}$  et le plan horizontal, qui est, l'angle compris entre  $\mathbf{M}_1\mathbf{a}$  et  $\mathbf{s}_2$ . La définition habituelle du cosinus d'un angle nous dit alors que  $\langle \mathbf{s}_2, \mathbf{M}_1\mathbf{a} \rangle$ , le produit intérieur de  $\mathbf{s}_2$  et  $\mathbf{M}_1\mathbf{a}$ , est

$$\langle \mathbf{s}_2, \mathbf{M}_1\mathbf{a} \rangle = \|\mathbf{s}_2\| \|\mathbf{M}_1\mathbf{a}\| \cos \phi. \quad (13.33)$$

Si nous écrivons  $\mathbf{a} = \mathbf{M}_1\mathbf{a} + \mathbf{P}_1\mathbf{a}$ , nous voyons que  $\langle \mathbf{s}_2, \mathbf{M}_1\mathbf{a} \rangle = \langle \mathbf{s}_2, \mathbf{a} \rangle$ , parce que  $\mathbf{s}_1$  est orthogonal à  $\mathbf{s}_2$ . Par (13.29),  $\langle \mathbf{s}_2, \mathbf{a} \rangle = c_2$ . Si nous rappelons plus loin que  $\mathbf{s}_2$  est un vecteur unitaire, tel que  $\|\mathbf{s}_2\| = 1$ , (13.33) deviennent

$$c_2 = \|\mathbf{M}_1\mathbf{a}\| \cos \phi.$$

Il s'en suit que le paramètre de non centralité  $\Lambda$ , que nous avons vu était égal à  $c_2^2$  dans le cas simple présent, est donné par une formule qui rappelle fortement

l'expression (12.23) pour le cas des tests dans des directions de régressions:

$$\Lambda = \|\mathbf{M}_1 \mathbf{a}\|^2 \cos^2 \phi. \quad (13.34)$$

Les arguments présentés au-dessous ne réclament pas de rigueur. En particulier, nous avons ignoré les distinctions entre les quantités calculées pour les échantillons de taille finie et les limites vers lesquelles ils tendent quand la taille de l'échantillon vers l'infini. En dépit de ces imperfections, notre discussion contient le coeur du sujet. La formule (13.34), bien que dérivée seulement pour un cas spécial, est en fait généralement valide, par laquelle nous signifions que non seulement que la formule correcte en général est la forme algébrique, mais aussi qu'il est propre de replacer les variables aléatoires utilisées en construisant la représentation géométrique dans un espace Euclidien par leurs limites en probabilité. Ainsi, le puissance des tests classiques est gouvernée par les mêmes considérations que les tests dans les directions de régression traitées dans le dernier chapitre. Cela dépend de deux choses: la distance entre le DGP et l'hypothèse nulle, lorsque mesuré par  $\|\mathbf{M}_1 \mathbf{a}\|$ , et l'"angle" compris entre le vecteur  $\mathbf{M}_1 \mathbf{a}$ , mesurant le degré d'incorrection de l'hypothèse nulle, et le sous-espace  $\mathbf{s}_2$  engendré par les directions correspondant aux variations des paramètres de l'hypothèse (alternative). L'intuition est identique à ce qui fut présenté dans le dernier chapitre. Un traitement entièrement mathématique est, cependant, au-delà de la portée de ce livre. Les lecteurs intéressés sont rapportés à Davidson et MacKinnon (1987) et à certaines références apparentées citées à la Section 13.8.

## 13.4 TESTS CLASSIQUES ET RÉGRESSIONS LINÉAIRES

Nous avons vu à la Section 8.10 que les estimations des paramètres de la fonction de régression dans le modèle de régression linéaire sont identiques aux estimations NLS si on fait l'hypothèse que les aléas sont distribués normalement. A fortiori, ce résultat est aussi vrai pour les modèles de régression linéaire. Il est par conséquent intéressant de comparer les statistiques de test  $t$  et  $F$  pour tester les restrictions linéaires sur les modèles de régression linéaire, pour lesquels sous les conditions classiques les distributions exactes d'échantillon fini sont connues, avec les trois statistiques de test classiques, pour lesquels en général seule la distribution asymptotique est connue. Il en sort que nous pouvons en dire beaucoup plus concernant les relations parmi les trois tests classiques lorsque nous restreignons notre attention aux restrictions linéaires sur des modèles linéaires.

Les modèles contraints et non contraints, rencontrés en premier comme (3.18) et (3.19), sont

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{u} \quad \text{and} \quad (13.35)$$

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{u}, \quad (13.36)$$

où  $\mathbf{X}_1$  est  $n \times (k - r)$  et  $\mathbf{X}_2$  est  $n \times r$ . Comme nous nous intéressons à l'estimation ML, nous supposons maintenant que  $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . L'hypothèse maintenue est que (13.36) est vraie, et l'hypothèse nulle est que  $\beta_2 = \mathbf{0}$ . Le test standard  $F$ , ou le test  $t$  dans le cas où  $r = 1$ , était discuté à la Section 3.5 et il sera utilisé comme comparaison de base avec les trois statistiques de test classiques. Dans la notation du Chapitre 3, la statistique  $F$  peut être écrite comme

$$F = \frac{n - k}{r} \times \frac{\mathbf{y}^\top \mathbf{P}_{M_1 X_2} \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}}, \quad (13.37)$$

où nous avons utilisé (3.21), (3.30), et (3.32) pour obtenir cette expression. Nous comparons maintenant les statistiques LM, LR, et Wald statistics avec cette statistique  $F$ .

Asymptotiquement,  $r$  fois la statistique  $F$  (13.37) est distribuée comme une  $\chi^2(r)$  sous l'hypothèse nulle. En fait, comme nous le motrerons bientôt, elle tend vers la même variable aléatoire comme les trois statistiques de test classiques. Ceci est entièrement dû à la présence du paramètre  $\sigma$  dans le modèle de régression linéaire qui n'est pas l'égalité parfaite de  $rF$  et des trois tests classiques. Supposons que  $\sigma$ , à la place d'être un paramètre connu devant être estimé, soit en fait connu. Alors, pour le cas dans lequel  $\sigma = 1$  (une restriction sans importance, car si nous connaissons  $\sigma$  nous pourrions toujours renormaliser les données), la fonction de logvraisemblance pour le modèle (13.36) serait

$$\ell(\mathbf{y}, \beta) = -\frac{n}{2} \log 2\pi - \frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

Pour des données d'échantillon fournies  $\mathbf{y}$  et  $\mathbf{X}$ , ceci correspond exactement à la fonction quadratique du vecteur  $\beta$ . Les résultats de la Section 13.2 sont alors directement applicables, et il est facile de calculer les trois statistiques et de montrer qu'elles sont toutes égales à  $rF$ .

Nous retournons maintenant au cas plus intéressant dans lequel  $\sigma$  doit être estimé. Commençons par la statistique LR. Il est commode d'exprimer cette statistique en termes de la fonction de logvraisemblance concentrée (8.82). Pour le modèle non contraint (13.36), cette fonction de logvraisemblance concentrée est

$$\hat{\ell} \equiv -\frac{n}{2} \log((\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})) = -\frac{n}{2} \log(\mathbf{y}^\top \mathbf{M}_X \mathbf{y}),$$

à part un terme constant qui est le même pour l'estimation à la fois de (13.35) et (13.36) et qui disparaît donc de la différence des fonctions de logvraisemblance utilisées dans le test LR. Ici  $\mathbf{X} \equiv [\mathbf{X}_1 \quad \mathbf{X}_2]$  et  $\mathbf{M}_X$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X})$ . Pour le modèle (13.35), la fonction de logvraisemblance concentrée est

$$\tilde{\ell} \equiv -\frac{n}{2} \log((\mathbf{y} - \mathbf{X}_1 \tilde{\beta}_1)^\top (\mathbf{y} - \mathbf{X}_1 \tilde{\beta}_1)) = -\frac{n}{2} \log(\mathbf{y}^\top \mathbf{M}_1 \mathbf{y}),$$

où  $\mathbf{M}_1$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}_1)$ . Ainsi, la statistique LR est

$$LR = 2(\hat{\ell} - \tilde{\ell}) = n \log \left( \frac{\mathbf{y}^\top \mathbf{M}_1 \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}} \right). \quad (13.38)$$

Il est facile de montrer que

$$\mathbf{y}^\top \mathbf{M}_1 \mathbf{y} = \mathbf{y}^\top \mathbf{M}_X \mathbf{y} + \mathbf{y}^\top \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y}.$$

Cette décomposition, qui a été expliquée dans la figure Figure 1.7, dit que la SSR d'une régression de  $\mathbf{y}$  sur  $\mathbf{X}_1$  est égale à la SSR d'une régression de  $\mathbf{y}$  sur  $\mathbf{X}_1$  et  $\mathbf{X}_2$ , plus la somme expliquée des carrés d'une régression de  $\mathbf{y}$  (ou, de manière équivalente,  $\mathbf{M}_1 \mathbf{y}$ ) sur  $\mathbf{M}_1 \mathbf{X}_2$ . En conséquence, nous obtenons de (13.38) que

$$LR = n \log \left( 1 + \frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}} \right). \quad (13.39)$$

La relation entre la statistique  $F$  (13.37) et la statistique LR (13.39) est alors

$$LR = n \log \left( 1 + \frac{rF}{n-k} \right). \quad (13.40)$$

Pour  $n$  grand, *pourvu* que  $F = O(1)$ , nous étendre Taylor au logarithme. Il s'agit clairement du cas de l'hypothèse nulle (13.35) et aussi, en fait, de celui des DGP qui se mettent en marche pour l'hypothèse nulle à un taux proportionnel à  $n^{-1/2}$ . La dernière assertion est facilement démontrée dans le cas d'un DGP qui se met en marche dans une direction de régression, comme (12.06), et peut avec quelque effort être montré comme étant vrai pour des courants des mises en marche de DGP, telles que (13.27). Le résultat du développement de Taylor est

$$LR = \left( \frac{n}{n-k} \right) rF + O(n^{-1}) = rF + O(n^{-1}),$$

qui démontre que  $LR$  et  $rF$  constituent la même variable aléatoire asymptotiquement.

Nous considérons ensuite la statistique de Wald,  $W$ . Pour les modèles (13.35) et (13.36) c'est, par (13.05) et (13.13),

$$W = \hat{\beta}_2^\top (\hat{\mathbf{I}}^{-1})_{22}^{-1} \hat{\beta}_2. \quad (13.41)$$

Pour le modèle de régression linéaire (13.36), nous avons de (8.85) qui le bloc  $(\beta, \beta)$  de  $\mathbf{I}$ , qui est tout nous avons besoin donné par la propriété bloque-diagonale de (8.87), est donné par ? à revoir?

$$(\mathbf{I}_{\beta\beta})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Naturellement,  $\sigma^2$  doit être estimée; comme nous sommes dans le contexte du maximum de vraisemblance, l'utilisation de l'estimateur ML est sensée

$$\hat{\sigma}^2 = \frac{1}{n} \mathbf{y}^\top \mathbf{M}_X \mathbf{y}.$$

Par le théorème FWL,

$$\begin{aligned} \hat{\beta}_2 &= (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y} \quad \text{and} \\ ((\mathbf{X}^\top \mathbf{X})^{-1})_{22} &= (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1}. \end{aligned}$$

Ainsi, (13.41) devient

$$W = n \left( \frac{\mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}} \right) = n \left( \frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_X \mathbf{y}} \right).$$

de (13.37) et (13.39), nous obtenons

$$W = \left( \frac{rn}{n-k} \right) F; \quad LR = n \log \left( 1 + \frac{W}{n} \right). \quad (13.42)$$

Comme  $W$  est égale à  $n/(n-k)$  fois  $rF$ , il est évident que

$$W = rF + O(n^{-1}).$$

Pour terminer, nous nous tournons vers la statistique LM. Nous observons tout d'abord de (8.83) que le gradient par rapport aux paramètres de régression  $\beta$  de la fonction de vraisemblance pour un modèle de régression linéaire à erreurs normales est

$$\mathbf{g}(\mathbf{y}, \beta, \sigma) = \frac{1}{\sigma^2} \sum_{t=1}^n \mathbf{X}_t^\top (y_t - \mathbf{X}_t \beta) = \sigma^{-2} \mathbf{X}^\top (\mathbf{y} - \mathbf{X} \beta).$$

Ainsi, de (13.03), la statistique LM est

$$\begin{aligned} LM &= \tilde{\mathbf{g}}_2^\top (\tilde{\mathbf{I}}^{-1})_{22} \tilde{\mathbf{g}}_2 \\ &= \tilde{\sigma}^{-4} (\mathbf{y} - \mathbf{X} \tilde{\beta})^\top \mathbf{X}_2 (\tilde{\sigma}^2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1}) \mathbf{X}_2^\top (\mathbf{y} - \mathbf{X} \tilde{\beta}). \end{aligned} \quad (13.43)$$

Comme le test LM est basé sur le modèle contraint (13.35), nous utilisons l'estimation ML de  $\sigma$  provenant de ce modèle:

$$\tilde{\sigma}^2 = \frac{1}{n} \mathbf{y}^\top \mathbf{M}_1 \mathbf{y}.$$

En substituant ceci dans (13.43), nous voyons que

$$\begin{aligned}
 LM &= n \left( \frac{\mathbf{y}^\top \mathbf{M}_1 \mathbf{X}_2 (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{M}_1 \mathbf{y}}{\mathbf{y}^\top \mathbf{M}_1 \mathbf{y}} \right) \\
 &= n \left( \frac{\mathbf{y}^\top \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y} / \mathbf{y}^\top \mathbf{M}_X \mathbf{y}}{1 + \mathbf{y}^\top \mathbf{P}_{\mathbf{M}_1 \mathbf{X}_2} \mathbf{y} / \mathbf{y}^\top \mathbf{M}_X \mathbf{y}} \right) \\
 &= n \left( \frac{rF}{n - k + rF} \right).
 \end{aligned} \tag{13.44}$$

Pour  $n$  grand, le développement de Taylor donne

$$LM = rF + O(n^{-1}).$$

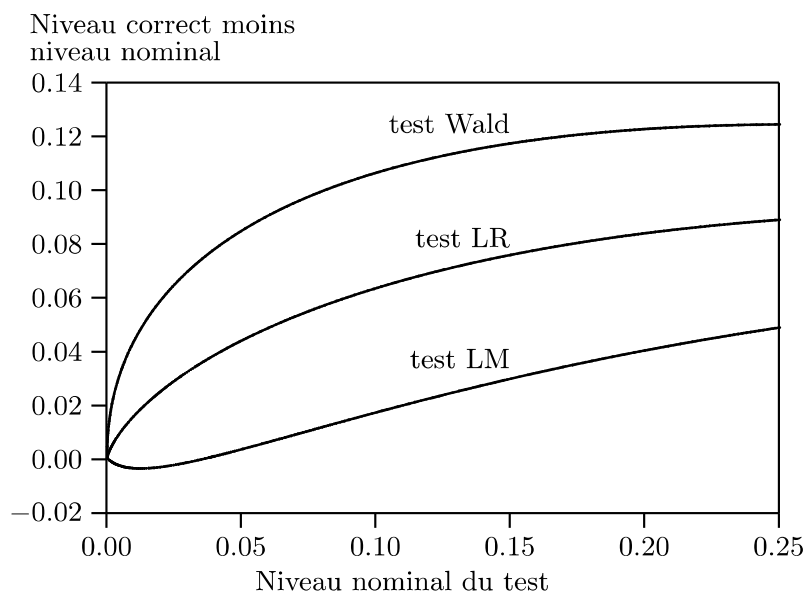
La conclusion la plus importante de cette analyse est que toutes ces statistiques de test classiques sont *seulement* des fonctions de la statistique standard  $F$  et des entiers  $n$ ,  $k$ , et  $r$ . Ainsi, si l'on calcule  $F$ , les statistiques  $LM$ ,  $W$ , et  $LR$  peuvent être obtenues directement de (13.44), (13.42), et (13.40). Si les conditions de régularité tiennent, telles que l'hypothèse nulle  $F$  ait exactement sa distribution de même nom, les distributions d'échantillon fini *exactes* de  $LM$ ,  $W$ , et  $LR$  peuvent être calculées par l'utilisation de ces formules. Cependant, ces distributions exactes *ne sont pas* les mêmes que la distribution asymptotique, qui est chi-carré (centrale) à  $r$  degrés de liberté.

Pour les modèles de régression linéaire, avec ou sans erreurs nomales, il existe naturellement aucune nécessité de regarder  $LM$ ,  $W$ , et  $LR$ , car l'information qui est obtenues en procédant ainsi en plus, est déjà contenue dans  $F$ . Néanmoins, comme les travailleurs appliqués trouvent souvent comme d'utiliser une de ces statistiques classiques de test, cela vaut la peine de discuter d'un problème qui peut survenir lorsqu'elles sont utilisées. Chaque des statistiques de test classiques sera typiquement comparée, pour les cas d'inférence, avec la distribution asymptotique chi-deux. Comme les trois sont numériquement différentes entre elles, différentes inférences peuvent bien être dessinées lorsque différents tests classiques sont employés. Cette difficulté, souvent rattachée à un **conflit parmi les différents critères de test**, est fréquemment composé par des manières diverses de calculer même juste une des statistiques classiques, comme nous en discuterons dans la prochaine section. Le résultat des conflits parmi les différents critères de test a bien été discuté dans la littérature économétriques. Le propos semble avoir été évoqué par Savin (1976) et Berndt et Savin (1977); il fut exposé et étendu par Breusch (1979). Consulter aussi Evans et Savin (1982) et Vandaele (1981).

Pour le cas des modèles de régression linéaire (incluant la GNR et les régressions artificielles pour ML que nous introduirons bientôt), et en de manière quelque peu générale, existent des relations d'inégalité qui tiennent parmi  $LM$ ,  $W$ , et  $LR$ . Ces inégalités sont comme suit:

$$W > LR > LM.$$





**Figure 13.5** Différences entre les niveaux corrects et nominaux

Ce résultat provient directement de (13.40), (13.42), et (13.44), avec les les inégalités suivantes standards pour  $x > 0$ :

$$x > \log(1+x) > \frac{x}{1+x}. \quad (13.45)$$

Ces inégalités standards sont faciles à démontrer. La première provient du résultat que

$$e^x = 1 + x + \frac{\delta x^2}{2!} > 1 + x, \quad (13.46)$$

pour n'importe quel  $\delta$  compris entre 0 et 1. L'égalité, ici, est une conséquence du Théorème de Taylor. En prenant le logarithme des deux cotés, (13.46) donne alors la première inégalité de (13.45). Pour la seconde inégalité, on peut remplacer  $x$  par  $-y$  dans (13.46) pour obtenir le résultat que  $e^{-y} > 1 - y$ . La prise des logarithmes donne

$$-\log(1-y) > y. \quad (13.47)$$

En mettant  $y = x/(1+x)$  dans (13.47), cela donne alors l'inégalité désirée.

On commet inévitablement une erreur si l'on compare une des statistiques de test classiques avec sa distribution asymptotique nominale dans un échantillon fini. Comme exemple, nous montrons dans la Figure 13.5 un graphique de taille réelle, calculé à partir des points de la distribution asymptotique  $F(1, 25)$ , des tests qui utilisent  $LM$ ,  $LR$ , et  $W$  comme une fonction de taille nominale donnée par la distribution asymptotique  $\chi^2(1)$ . L'inégalité  $W > LR > LM$  apparaît très évidente sur la figure, et à la fois  $LR$  et  $W$  sont vues ?surrejeter? très sévèrement.

Tous les résultats présentés dans cette section semblent jusqu'ici conduire raisonnablement à une conclusion claire. Chaque fois qu'une hypothèse de test est appliquée par l'utilisation d'une régression linéaire, la forme la plus facile et souvent la plus satisfaisante de la statistique de test pour l'utilisation est le test  $F$  ou, dans les cas à une dimension, le test  $t$ . Toutes les autres statistiques de test que nous avons considérées jusqu'ici sont fonction de la statistique  $F$  ou  $t$ , sont asymptotiquement équivalents sous l'hypothèse nulle et sous les DGP qui se mettent en marche pour l'hypothèse nulle à un taux proportionnel à  $n^{-1/2}$ , mais peuvent avoir dans des échantillons finis des distributions perturbatrices loin de la distribution asymptotique nominale.

## 13.5 AUTRES ESTIMATEURS DE LA MATRICE DE COVARIANCE

Pour calculer à la fois les tests LM et Wald tests, on a à employer un certain estimateur de la matrice d'information,  $\mathcal{J}_0$ . Jusqu'ici, nous avons supposé, au moins implicitement, que la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$  est connue comme une fonction du vecteur paramétrique et est alors simplement évalué en  $\hat{\boldsymbol{\theta}}$  dans le cas du test de Wald ou  $\tilde{\boldsymbol{\theta}}$  dans le cas du test de LM. Ceci est certainement vrai pour les modèles de régression linéaire, où la matrice d'information ne dépend même pas des paramètres de la fonction de régression. Dans ces cas, comme nous l'avons vu dans la dernière section, seul le choix de  $\hat{\sigma}^2$  ou de  $\tilde{\sigma}^2$  distingue le test de Wald de celui de LM.

En général, cependant, comme nous l'avons vu dans la Section 8.6, il n'est pas réaliste de supposer que la matrice d'information soit connue sous une forme analytique explicite. Quand cela n'est pas le cas, il devient nécessaire d'utiliser un de ce qui peut être une large variété d'estimateurs de  $\mathcal{J}_0$ . Pourvu que l'estimateur choisi soit convergent, aucun des résultats asymptotiques établis jusqu'ici n'est affecté par ce choix. Cependant, le comportement d'échantillon fini des tests peut bien dépendre de l'estimateur qui est utilisé. Lorsque différentes variantes du test LM ou Wald, basées sur différents estimateurs de la matrice d'information, sont utilisés dans des échantillons finis, il existe la possibilité que les résultats de test peuvent être incompatibles. Naturellement, ce problème ne survient pas avec le test LR, car il n'emploie pas un estimateur de la matrice d'information.

Une autre source possible de contradiction parmi les différents tests apparaît lorsque nous considérons comment les tests classiques se comportent sous des reparamétrisations de l'hypothèse nulle ou alternative. Nous regarderons les reparamétrisations plus en détail dans la prochaine section. Dans cette section, cependant, nous verrons que, même si nous décidons d'un estimateur de la matrice d'information, les tests LM et Wald ne sont pas invariants sous une reparamétrisation modèle pour tous estimateurs de la sorte.

Nous voulons illustrer maintenant ces problèmes dans le contexte d'un exemple, qui, bien que très simple, exhibe la plupart des questions pendantes.

L'exemple concerne un modèle dans lequel on souhaite estimer la variance  $\sigma^2$  d'une série de variables aléatoires n.i.d. avec moyenne *connue* qui peut, sans perte de généralité, être prise comme zéro. La contribution de la fonction logvraisemblance d'une observation est

$$\ell_t(y_t, \sigma^2) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{y_t^2}{2\sigma^2},$$

et la fonction de logvraisemblance pour un échantillon de taille  $n$  est

$$\ell(\mathbf{y}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^n y_t^2. \quad (13.48)$$

Ce modèle n'a juste qu'un paramètre, qui devrait normalement être pris comme soit  $\sigma$ , soit  $\sigma^2$ . Cependant, il est intéressant de considérer une troisième paramétrisation. Supposons que  $\tau \equiv \log \sigma$  soit le paramètre du modèle. Alors la fonction de logvraisemblance devient

$$\ell(\mathbf{y}, \tau) = -\frac{n}{2} \log(2\pi) - n\tau - \frac{1}{2} e^{-2\tau} \sum_{t=1}^n y_t^2, \quad (13.49)$$

de laquelle nous pouvons tirer que

$$e^{2\hat{\tau}} = \hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n y_t^2. \quad (13.50)$$

Pour cette reparamétrisation, la matrice d'information, qui possède seulement un élément, est constante et égale à 2:

$$\mathcal{I} = -\frac{1}{n} E(D_\tau^2 \ell) = \frac{2}{n} \sum_{t=1}^n e^{-2\tau} E(y_t^2) = 2.$$

Notons que, bien que  $\mathcal{I}$  soit constante, la fonction de logvraisemblance *n'est pas* une fonction quadratique de  $\tau$ . Nous considérons maintenant des tests classiques variés pour l'hypothèse nulle que  $\tau = 0$ , ou, de manière équivalente, que  $\sigma^2 = 1$ . En dépit de la simplicité de cet exemple, nous dévoilerons une variété ahurissante de statistiques de test.

Initialement, nous travaillerons avec la paramétrisation  $\tau$ . Il n'est pas nécessaire du tout de procéder à une quelconque estimation afin de trouver des estimations contraintes, car  $\tilde{\tau} = 0$ . Pour les tests Wald et LR, nous avons besoin de trouver  $\hat{\tau}$ . De (13.50), c'est

$$\hat{\tau} = \frac{1}{2} \log \left( \frac{1}{n} \sum_{t=1}^n y_t^2 \right).$$

Le “maximum” contraint de la fonction de logvraisemblance est just la valeur de la fonction en  $\tau = 0$ :

$$\tilde{\ell} = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^n y_t^2 = -\frac{n}{2} \log 2\pi - \frac{n}{2} e^{2\hat{\tau}}. \quad (13.51)$$

Bien que ceci soit le maximum contraint, il est commode de l’exprimer, comme nous l’avons fait ici, en termes d’une estimation non contrainte,  $\hat{\tau}$ . Le maximum non contraint,  $\hat{\ell}$ , est donné par

$$-\frac{n}{2} \log 2\pi - n\hat{\tau} - \frac{1}{2} e^{-2\hat{\tau}} \sum_{t=1}^n y_t^2 = -\frac{n}{2} \log 2\pi - n\hat{\tau} - \frac{n}{2}, \quad (13.52)$$

où l’égalité utilise (13.50).

Nous pouvons poursuivre immédiatement pour obtenir la statistique LR, qui est deux fois la différence entre (13.52) et (13.51):

$$\begin{aligned} LR = 2(\hat{\ell} - \tilde{\ell}) &= n(e^{2\hat{\tau}} - 1 - 2\hat{\tau}) \\ &= 2n\hat{\tau}^2 + o(1). \end{aligned} \quad (13.53)$$

La seconde ligne de (13.53) est un développement de Taylor de la statistique dans des puissances de  $\hat{\tau}$ . Ceci est intéressant parce que, *sous l’hypothèse nulle*, nous attendons  $\hat{\tau}$ , qui est à la fois l’estimation elle-même et la *différence* entre l’estimation et la véritable valeur du paramètre, pour être de l’ordre de  $n^{-1/2}$ . Il s’en suit que  $2n\hat{\tau}^2$  sera de l’ordre de l’unité et que les termes plus hauts dans le développement de la fonction exponentielle dans (13.53) sera d’un ordre plus bas. Ainsi, si les formes variées du test classique donne en effet des expressions asymptotiquement égales, nous pouvons nous attendre à ce que le terme induit d’entre elles sera  $2n\hat{\tau}^2$ .

Considérons ensuite la statistique LM. La pièce essentielle de celle-ci est la dérivée de la fonction de logvraisemblance (13.49) par rapport à  $\tau$ , évaluée en  $\tau = 0$ . Nous trouvons que

$$\frac{\partial \ell}{\partial \tau} = -n + e^{-2\tau} \sum_{t=1}^n y_t^2 \quad \text{and} \quad \left. \frac{\partial \ell}{\partial \tau} \right|_{\tau=0} = n(e^{2\hat{\tau}} - 1). \quad (13.54)$$

Si pour la variance de  $\partial \ell / \partial \tau$  nous utilisons  $n$  fois la véritable valeur constante, du seul élément de la matrice d’information, 2, la statistique LM est le carré de  $(\partial \ell / \partial \tau)|_{\tau=0}$ , given by (13.54), divisé par  $2n$ :

$$LM_1 = \frac{n}{2} (e^{2\hat{\tau}} - 1)^2 = 2n\hat{\tau}^2 + o(1).$$

Cette variante de la statistique LM possède le même terme induit que la statistique LR (13.53) mais différera naturellement d’elle dans les échantillons finis.

A la place de la véritable matrice d'information, un enquêteur pourrait préférer utiliser la négative du Hessien empirique pour estimer la matrice d'information; consulter les équations (8.47) et (8.49). Comme la fonction de logvraisemblance n'est pas exactement quadratique, cette estimateur *ne coïncide pas* numériquement avec la véritable valeur. Comme

$$\frac{\partial^2 \ell}{\partial \tau^2} = -2e^{-2\tau} \sum_{t=1}^n y_t^2, \quad (13.55)$$

qui en  $\tau = 0$  est  $-2ne^{2\hat{\tau}}$ , le test LM calculé de cette manière

$$LM_2 = \frac{n}{2} e^{-2\hat{\tau}} (e^{2\hat{\tau}} - 1)^2 = 2n\hat{\tau}^2 + o(1). \quad (13.56)$$

Le terme induit est comme dans  $LR$  et  $LM_1$ , mais  $LM_2$  différera de ces deux statistiques dans des échantillons finis.

Une autre possibilité consiste à utiliser l'estimateur OPG de la matrice d'information; consulter les équations (8.48) et (8.50). Cet estimateur est

$$\frac{1}{n} \sum_{t=1}^n \left( \frac{\partial \ell}{\partial \tau} \right)^2 = \frac{1}{n} \sum_{t=1}^n (y_t^2 e^{-2\tau} - 1)^2,$$

qui, lorsqu'évaluée en  $\tau = 0$ , est égale à

$$\frac{1}{n} \sum_{t=1}^n (y_t^2 - 1)^2.$$

Cette expression ne peut même pas être exprimée comme une seule fonction de  $\hat{\tau}$ . Pour obtenir un développement de la statistique de test qui l'utilise, nous devons employer la propriété de la distribution normale qui nous dit que  $E(y_t^4) = 3\sigma^4$ , ou, en termes de  $\tau$ ,  $3e^{4\tau}$ .<sup>4</sup> En utilisant cette propriété, nous pouvons invoquer une loi des grands nombre et conclure que l'estimateur de la matrice d'information OPG est en effet égale à  $2 + o(1)$  at  $\tau = 0$ . Ainsi, la troisième variante de la statistique LM est

$$LM_3 = \frac{n^2 (e^{2\hat{\tau}} - 1)^2}{\sum_{t=1}^n (y_t^2 - 1)^2} = 2n\hat{\tau}^2 + o(1).$$

<sup>4</sup> Notons que cela *ne fut pas* nécessaire pour utiliser les propriétés spéciales de la distribution afin de les statistiques antérieures, qui étaient en fait toutes des fonctions d'une et esulement une variable aléatoire, à savoir  $\hat{\tau}$ . En général, dans les situations les moins simples, cette caractéristique agréable du présent exemple est absente et les propriétés doivent être invoquées afin de découvrir le comportement de toutes les statistiques de test variées.

Une fois encore, le terme induit est  $2n\hat{\tau}^2$ , mais la forme de  $LM_3$  est autrement très différente de celle de  $LM_1$  ou  $LM_2$ .

Tous comme il existent des formes variées du test LM, il existe également des formes variées du test de Wald. N'importe laquelle d'entre elles peut être formée par la combinaison de l'estimation non contrainte  $\hat{\tau}$  avec une quelconque estimation de la matrice d'information, qui dans ce cas est effectivement un scalaire. Le choix le plus simple est juste la véritable matrice d'information, qui est, 2. Avec ceci nous obtenons

$$W_1 = 2n\hat{\tau}^2. \quad (13.57)$$

Il est facile de voir que  $W_2$ , qui utilise le Hessien empirique, est identique à  $W_1$ , parce (13.55) évalué en  $\tau = \hat{\tau}$  est juste  $2n$ . D'un autre côté, l'utilisation de l'estimateur OPG donne

$$W_3 = \hat{\tau}^2 \sum_{t=1}^n (y_t^2 e^{-2\hat{\tau}} - 1)^2,$$

qui est très différente de  $W_1$  et  $W_2$ .

Toutes les statistiques de test du dessus étaient basées sur  $\tau$  comme l'unique du modèle, mais nous pourrions tout aussi bien utiliser  $\sigma$  ou  $\sigma^2$  comme modèle paramétrique. L'idéal pour nous serait que les statistiques de test soient invariantes à de telles reparamétrisations. La statistique LR est toujours invariante, parce que  $\hat{\ell}$  et  $\tilde{\ell}$  demeurent inchangées lorsque le modèle est reparamétrisé. Mais toutes les formes de la statistique de Wald, et certaines formes de la statistique LM, ne sont pas en général invariantes, comme nous l'illustrons à présent.

Supposons que nous prenions  $\sigma^2$  pour être le paramètre du modèle. La matrice d'information n'est pas constante dans cette nouvelle paramétrisation, et ainsi nous devons l'évaluer en l'estimation  $\hat{\sigma}^2$ . Il est facile de voir que la matrice d'information, comme une fonction de  $\sigma^2$ , est  $1/(2\sigma^4)$ . Si nous utilisons cette expression pour la matrice d'information, évaluée en  $\hat{\sigma}^2$ , le test de Wald

$$W_1 = \frac{n}{2} \hat{\sigma}^{-4} (\hat{\sigma}^2 - 1)^2 = \frac{n}{2} e^{-4\hat{\tau}} (e^{2\hat{\tau}} - 1)^2 = 2n\hat{\tau}^2 + o(1).$$

Comme celle-ci diffère de (13.57), nous devons montrer que les différentes paramétrisations mènent à des statistiques de Wald numériquement différentes même si la véritable matrice d'information, évaluées au MLE du paramètre du modèle, est utilisées dans les deux cas.

Comme nous le verrons dans la prochaine section, le test LM est invariant si il est basé sur la véritable matrice d'information évaluée en l'MLE. Mais si un autre estimateur de la matrice d'information est utilisé, le test LM peut aussi être dépendant de la paramétrisation. Supposons que nous utilisons le

Hessien empirique. De (13.48), les deux dérivées premières de  $\ell$  par rapport à  $\sigma^2$ , évaluées en  $\sigma^2 = 1$ , sont

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^2} \Big|_{\sigma^2=1} &= -\frac{1}{2} \left( n - \sum_{t=1}^n y_t^2 \right) = \frac{n}{2} (e^{2\hat{\tau}} - 1) \quad \text{and} \\ \frac{\partial^2 \ell}{(\partial \sigma^2)^2} \Big|_{\sigma^2=1} &= \frac{n}{2} (1 - 2e^{2\hat{\tau}}).\end{aligned}$$

De ceci, nous trouvons que la statistique  $LM_2$  calculée comme le fut (13.56) mais pour la paramétrisation  $\sigma^2$ , est

$$LM_2 = \frac{n(e^{2\hat{\tau}} - 1)^2}{2(2e^{2\hat{\tau}} - 1)} = 2n\hat{\tau}^2 + o(1). \quad (13.58)$$

Le terme dominant est correct, comme il doit l'être, mais (13.58) est numériquement différente de (13.56). ■

De toute évidence, il existe encore plus de formes des deux tests LM et Wald, mais elles ne coïncideront pas toutes avec une des versions que nous avons déjà calculées. Le lecteur intéressé est invité à expérimenter, par exemple, les effets de l'utilisation  $\sigma$  lui-même, plutôt que de  $\sigma^2$  comme paramètre du modèle.

Cet exemple illustre le fait qu'il peut exister différentes statistiques de test classiques, qui sont numériquement différentes mais asymptotiquement équivalentes. Le fait qu'il existe tant de tests différents crée le problème de comment établir un coix parmi eux. On préférerait utiliser les tests qui sont faciles à calculer et pour lesquels la distribution d'échantillon fini est bien approximée par la distribution asymptotique. Malheureusement, cela demande fréquemment des efforts considérables pour déterminer les propriétés d'échantillons finis des tests asymptotiques. N'importe quelle méthode d'analyse tend à être contrainte à des cas très spéciaux, tels que le cas des modèles de régression à erreurs normales discutés à la Section 13.4. Une approche généralement applicable est d'employer la simulation informatique (les expériences Monte Carlo); consulter le Chapitre 21.

## 13.6 LES STATISTIQUES DE TEST ET LA REPARAMÉTRISATION

L'idée d'une **reparamétrisation** d'un modèle paramétrisé a été discuté en long dans la Section 8.3. Nous y avons vu qu'une des propriétés de l'estimation par maximum de vraisemblance est son invariance sous des reparamétrisations. Comme les statistiques de test sont comprises dans le contexte de l'estimation par maximum de vraisemblance, il pourrait être attendu, ou du moins espérer, que les statistiques de test classiques soient également invariantes à la paramétrisation. Ceci est vrai pour la statistique LR statistic, car comme

cela fut montré au Chapitre 8, la valeur d'une fonction de logvraisemblance maximisée est invariante à la reparamétrisation. Mais les résultats de la dernière section ont montré qu'il ne peut pas être vrai en général pour les autres tests classiques. Dans cette section, nous discutons des effets de la reparamétrisation des statistiques de test classiques plus en détail. En particulier, nous nous efforçons de déterminer que les éléments des tests LM et Wald, et que les éléments des estimateurs de matrice d'information, sont ou ne sont pas responsables pour la dépendance de nombreuses autres formes possibles des tests classiques. Nous croyons que ceux-ci constituent des préoccupations importantes. Cependant, la discussion est nécessairement très détaillée, et certains lecteurs peuvent souhaiter sauter cette section en première lecture.

Tout d'abord, nous devons expliquer que quand nous parlons de l'**invariance** nous signifions différentes choses lorsque nous discutons de différentes quantités. Par exemple, si un modèle est reparamétrisé par une application  $\eta : \Theta \rightarrow \Phi$ , où  $\theta$  et  $\phi$  désignent les vecteurs paramétriques sous les deux reparamétrisations, alors par l'invariance de l'MLE sous reparamétrisation, cela ne signifie certainement pas que  $\hat{\theta} = \hat{\phi}$ , mais plutôt que

$$\hat{\phi} = \eta(\hat{\theta}). \quad (13.59)$$

la notation ici a été utilisée au préalable au Chapitre 8, autour de l'équation (8.23), et sera utilisée encore au-dessous. Nous devons distinguer entre les quantités exprimées en termes de vecteur  $k$ -des paramètres  $\theta$  et les quantités exprimées en termes de vecteur  $k$ -des paramètres  $\phi$ . Comme dans le Chapitre 8, nous utiliserons les principales afin de signifier les quantités exprimées en termes de  $\phi$ .

Pour la fonction de logvraisemblance, maximisée, l'invariance signifie simplement que

$$\ell(\hat{\theta}) = \ell'(\hat{\phi}).$$

Ainsi, lorsque nous parlons des estimations paramétriques comme étant invariantes sous une reparamétrisation, nous voulons dire que (13.59) tient, tandis que lorsque nous parlons des fonctions de logvraisemblance maximisées, ou des statistiques de test, nous voulons dire que la réelle valeur numérique demeure inchangée quand elle est calculées en utilisant différentes paramétrisations.

Les tests Wald et LM sont confectionnés d'éléments qui sont des vecteurs et des matrices, contrairement au test LR qui dépend juste de deux scalaires. Afin de déterminer si oui ou non les quantités scalaires qui sont définies en termes de vecteurs et de matrices, telles que les statistiques de test classiques, sont invariantes, nous devons premièrement déterminer comment les vecteurs et les matrices eux-mêmes sont altérés par une reparamétrisation. Cela peut être mené par la suite si ces remaniements s'annulent dans la définition du scalaire. Des définitions (13.03) et (13.05) des tests LM et Wald, il peut être vu que les vecteurs et les matrices dont nous avons besoin pour examiner sont  $g(\theta)$ , le gradient de la fonction de logvraisemblance,  $\mathcal{J}(\theta)$ , la ma-



trice d'information,  $\mathbf{r}(\boldsymbol{\theta})$ , les contraintes de vecteurs, et  $\mathbf{R}(\boldsymbol{\theta})$ , la matrice des dérivées des éléments de  $\mathbf{r}(\boldsymbol{\theta})$ .

Nous considérerons une paramétrisation dans laquelle le “nouveau” vecteur paramétrique  $\boldsymbol{\phi}$  de dimension  $k$ -, est relié à l’“ancien” vecteur paramétrique  $\boldsymbol{\theta}$  de dimension  $k$ - par l'application  $\boldsymbol{\eta}$ :

$$\boldsymbol{\phi} = \boldsymbol{\eta}(\boldsymbol{\theta}).$$

Par conséquent, de ce qui suit immédiatement de (8.23), les fonctions de logvraisemblance dans les deux paramétrisations sont reliées par

$$\ell'(\mathbf{y}, \boldsymbol{\eta}(\boldsymbol{\theta})) = \ell(\mathbf{y}, \boldsymbol{\theta}). \quad (13.60)$$

Le vecteur gradient dans la paramétrisation d'origine  $\boldsymbol{\theta}$  est

$$\mathbf{g}(\boldsymbol{\theta}) \equiv D_{\boldsymbol{\theta}}^{\top} \ell(\mathbf{y}, \boldsymbol{\theta}), \quad (13.61)$$

et dans la paramétrisation  $\boldsymbol{\phi}$  c'est

$$\mathbf{g}'(\boldsymbol{\eta}(\boldsymbol{\theta})) \equiv D_{\boldsymbol{\phi}}^{\top} \ell'(\mathbf{y}, \boldsymbol{\eta}(\boldsymbol{\theta})). \quad (13.62)$$

La relation entre  $\mathbf{g}$  et  $\mathbf{g}'$  est obtenus par la différenciation l'identité définissant (13.60) par rapport aux éléments de  $\boldsymbol{\theta}$  et par l'utilisation la règle de série? :

$$D_{\boldsymbol{\phi}} \ell'(\boldsymbol{\eta}(\boldsymbol{\theta})) D_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta}) = D_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}), \quad (13.63)$$

où nous avons supprimés la dépendance sur  $\mathbf{y}$  par simplicité de notation. Notons que  $D_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})$  est une matrice  $k \times k$  avec comme élément type

$$\frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j}. \quad (13.64)$$

Désignons cette matrice, la matrice Jacobienne associée à la reparamétrisation  $\boldsymbol{\eta}$ , comme  $\mathbf{J}(\boldsymbol{\theta})$ . alors, de (13.63) et des définitions (13.61) et (13.62), nous obtenons

$$\mathbf{J}^{\top}(\boldsymbol{\theta}) \mathbf{g}'(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbf{g}(\boldsymbol{\theta}). \quad (13.65)$$

Ceci est le lien désiré entre les gradients dans les deux paramétrisations. Comme  $\boldsymbol{\eta}$  est une application inversible, il sera presque toujours vrai que sa Jacobienne  $\mathbf{J}(\boldsymbol{\theta})$  soit une matrice inversible (qui soit, non singulière) pour tout  $\boldsymbol{\theta} \in \Theta$ . Aussi si, nous pouvons alors inverser (13.65) de telle manière que  $\mathbf{g}'$  puisse être exprimé en termes de  $\mathbf{g}$ :

$$\mathbf{g}'(\boldsymbol{\eta}(\boldsymbol{\theta})) = (\mathbf{J}^{\top}(\boldsymbol{\theta}))^{-1} \mathbf{g}(\boldsymbol{\theta}). \quad (13.66)$$

Mais en général nous sommes obligés de supposer explicitement la non singularité de  $\mathbf{J}(\boldsymbol{\theta})$ , parce qu'il est possible de trouver par des reparamétrisations

$\boldsymbol{\eta}$  les Jacobiens qui sont singuliers pour certaines valeurs de  $\boldsymbol{\theta}$ .<sup>5</sup> Dans de tels cas, soit la reparamétrisation soit son inverse n'est pas lisse, dans le sens d'être continument différentiable. Ainsi, les résultats que nous mettons en réalisation en ce moment seront seulement vrais pour les reparamétrisations lisses qui ont des inverses lisses.

Pour la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$ , il est commode de commencer de sa définition, telle qu'exprimée dans l'équation (8.20), (8.21), et (8.22). De ces équations, nous pouvons conclure que la matrice d'information pour un exemple de taille  $n$  est

$$\mathbf{I}^n(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\mathbf{g}(\boldsymbol{\theta})\mathbf{g}^{\top}(\boldsymbol{\theta})). \quad (13.67)$$

Alors, de (13.65), nous trouvons que

$$E_{\boldsymbol{\theta}}(\mathbf{g}\mathbf{g}^{\top}) = \mathbf{J}^{\top} E_{\boldsymbol{\theta}}(\mathbf{g}'(\mathbf{g}')^{\top}) \mathbf{J},$$

où les quantités ? non amorcées?  $\mathbf{g}$  et  $\mathbf{J}$  sont évaluées en  $\boldsymbol{\theta}$ , et les quantités ?amorcées? sont évaluées en  $\boldsymbol{\phi}$ . Il s'en suit que

$$\mathbf{I}^n = \mathbf{J}^{\top} (\mathbf{I}^n)' \mathbf{J}.$$

La division par  $n$  et la prise de la limite quand  $n \rightarrow \infty$  donnent la règle de transformation pour la matrice d'information:

$$\mathcal{J} = \mathbf{J}^{\top} \mathcal{J}' \mathbf{J}. \quad (13.68)$$

Sous notre hypothèse de non singularité de  $\mathbf{J}$ , l'inverse de la règle de transformation est

$$\mathcal{J}' = (\mathbf{J}^{\top})^{-1} \mathcal{J} \mathbf{J}^{-1}. \quad (13.69)$$

Nous sommes maintenant prêts pour considérer la statistique LM sous la forme (13.03), qui est, la forme sous laquelle la matrice d'information correcte est usitée, évaluée en  $\tilde{\boldsymbol{\theta}}$ . Cette forme de la statistique de test est parfois appelée la statistique de test **score efficace**, par extension de la terminologie dans laquelle le test LM est appelé le test score (consulter la Section 8.9). Dans la paramétrisation  $\boldsymbol{\phi}$ , la forme score efficace du test LM devient

$$\frac{1}{n} (\mathbf{g}'(\tilde{\boldsymbol{\phi}}))^{\top} (\mathcal{J}'(\tilde{\boldsymbol{\phi}}))^{-1} \mathbf{g}'(\tilde{\boldsymbol{\phi}}), \quad (13.70)$$

où  $\tilde{\boldsymbol{\phi}} \equiv \boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})$  est, par l' "invariance" de l'MLE, l'MLE contraint dans la paramétrisation  $\boldsymbol{\phi}$ . Alors si, comme d'habitude,  $\tilde{\mathbf{g}}$  désigne  $\mathbf{g}(\tilde{\boldsymbol{\theta}})$  et ainsi de suite, (13.70) devient, par (13.66) et (13.69),

$$\frac{1}{n} \tilde{\mathbf{g}}^{\top} \tilde{\mathbf{J}}^{-1} \tilde{\mathbf{J}} \tilde{\mathbf{J}}^{-1} \tilde{\mathbf{J}}^{\top} (\tilde{\mathbf{J}}^{\top})^{-1} \tilde{\mathbf{g}} = \frac{1}{n} \tilde{\mathbf{g}}^{\top} \tilde{\mathbf{J}}^{-1} \tilde{\mathbf{g}},$$

<sup>5</sup> Par exemple, dans le cas du paramètre scalaire  $\theta$ , l'application inversible qui prend  $\theta$  dans  $\theta^3$  a un Jacobien de zéro en  $\theta = 0$ .

qui est juste la statistique (13.03) dans la paramétrisation d'origine  $\theta$ . Ainsi, nous pouvons conclure que la forme score efficace du test LM est en effet invariante sous les reparamétrisations, car tous les facteurs Jacobiens ont été enlevés de l'expression finale.

Nous considérons ensuite la statistique LM sous la forme sous laquelle la matrice d'information est estimée au moyen du Hessien Empirique, comme dans (13.56). Le Hessien empirique peut être écrit comme

$$\mathbf{H}(\boldsymbol{\theta}) \equiv D^2 \ell(\boldsymbol{\theta}) \quad (13.71)$$

dans la paramétrisation  $\theta$  et comme

$$\mathbf{H}'(\phi) \equiv D^2 \ell'(\phi) \quad (13.72)$$

dans la paramétrisation  $\phi$ . Si nous différencions (13.63) une fois de plus par rapport à  $\boldsymbol{\theta}$ , nous obtenons

$$D_{\boldsymbol{\theta}}^{\top} \boldsymbol{\eta}(\boldsymbol{\theta}) D_{\phi}^2 \ell'(\phi) D_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta}) + \sum_{i=1}^k \frac{\partial \ell'(\phi)}{\partial \phi_i} D_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 \boldsymbol{\eta}_i(\boldsymbol{\theta}) = D_{\boldsymbol{\theta}\boldsymbol{\theta}}^2 \ell(\boldsymbol{\theta}).$$

Avec le réarrangement et l'utilisation des définitions (13.71) et (13.72), ceci devient

$$\mathbf{H}(\boldsymbol{\theta}) = \mathbf{J}^{\top}(\boldsymbol{\theta}) \mathbf{H}'(\phi) \mathbf{J}(\boldsymbol{\theta}) + \sum_{i=1}^k \mathbf{g}'_i(\phi) D_{\boldsymbol{\theta}} \mathbf{J}_i(\boldsymbol{\theta}). \quad (13.73)$$

La notation utilisée pour le second terme du membre de droite au-dessus nécessite une petite exploitation. Tout d'abord, le terme entier doit être une matrice  $k \times k$  afin de s'accorder avec les autres termes dans l'équation; les  $\eta$  individue(les) dans le terme doivent alors être des matrices  $k \times k$  aussi. Ensuite,  $\mathbf{g}'_i(\phi)$  est juste le  $i$ ème élément du gradient  $\mathbf{g}'$ , évalué en  $\phi$ , et ainsi il est simplement un scalaire. Il s'en suit que  $D_{\boldsymbol{\theta}} \mathbf{J}_i(\boldsymbol{\theta})$  doit être une matrice  $k \times k$ . Si nous rappelons que  $\mathbf{J}(\boldsymbol{\theta})$  elle-même est une matrice  $k \times k$  avec comme élément type (13.64), nous voyons que  $\mathbf{J}_i(\boldsymbol{\theta})$ , la  $i$ ème ligne de la matrice, est  $1 \times k$ . Quand chacun des  $k$  éléments de la ligne est différencié par rapport aux  $k$  éléments de  $\boldsymbol{\theta}$ , nous obtenons pour finir la matrice  $D_{\boldsymbol{\theta}} \mathbf{J}_i(\boldsymbol{\theta})$  de dimension  $k \times k$  avec l'élément  $j$ l<sup>th</sup> donné par

$$\frac{\partial \mathbf{J}_{ij}(\boldsymbol{\theta})}{\partial \theta_l}. \quad (13.74)$$

Il peut maintenant être vu que la raison de la complexité notationnelle est qu'il existe trois indices variant de manière indépendante dans la dérivée partielle (13.74).

La relation (13.73) pour le Hessien devrait être parfaitement analogue à la relation (13.68) pour la matrice d'information si le second terme maladroite sur le coté droit de (13.73) était absents. Par conséquent, pour des

reparamétrisations qui sont telles que ce terme disparaît, la statistique LM calculée avec le Hessien empirique sera invariante. En général, cependant, ce terme ne disparaîtra pas et la statistique LM calculée avec le Hessien empirique ne sera pas invariante. Il existe des reparamétrisations pour lesquelles le terme maladroit dans (13.73) disparaît toujours, à savoir, **les reparamétrisations linéaires**. En effet, si  $\eta$  est une application linéaire, toutes les dérivées partielles du second ordre de la forme (13.74) sont nulles. Notons que les paramètres  $\tau$  et  $\sigma^2$  étudiés dans l'exemple de la section précédente *ne sont pas* liés par une relation linéaire.

Le **terme d'ordre induit** dans un développement asymptotique de la statistique LM doit naturellement être invariante pour la reparamétrisation par le résultat sur l'équivalence asymptotique. Le fait que ceci soit ainsi peut être vu en considérant les ordres de l'importance des trois termes dans (13.73). Parce que l'application  $\eta$  est indépendante de  $n$ , la matrice  $\mathbf{J}$  et les dérivées de ses éléments sont  $O(1)$ . Les Hessiens  $\mathbf{H}$  et  $\mathbf{H}'$  sont  $O(n)$ , et les gradients  $\mathbf{g}$  et  $\mathbf{g}'$  sont  $O(n^{1/2})$ . Nous voyons que le terme responsable pour la dépendance de paramétrisation de cette forme de la statistique LM n'est pas de l'ordre induit asymptotiquement, qui est  $O(n^{1/2})$ , tandis que l'autre terme du membre de droite de (13.73) est  $O(n)$ .

Il est clair que la possible dépendance de paramétrisation de la statistique LM est due seulement au choix de l'estimateur pour la matrice d'information. Ainsi, le choix d'un estimateur non invariant de matrice d'information tel que le Hessien empirique amènera la dépendance de paramétrisation dans une statistique de Wald tout comme dans une statistique LM. Cependant, dans l'exemple de la section précédente, nous avons vu que la statistique de Wald pouvait être dépendante de la reparamétrisation même si la matrice d'information réelle, évaluée au MLE non contraint, était utilisée. Ceci se révèle être une propriété générale du test de Wald: N'importe quelle reparamétrisation mènera à une valeur différente pour la statistique de test, sans se soucier de l'estimateur de la matrice d'information utilisé.

La non invariance du test de Wald a été le sujet de beaucoup de recherches. Des articles publiés par Gregory et Veall (1985, 1987) et Lafontaine et White (1986) ont mené à une étude plus détaillée par Phillips and Park (1988). Il apparaît que, pour une série de données fournie et une série de restrictions fournies sous une hypothèse non contrainte donnée, il est possible commodité de choix de paramétrisation d'obtenir *n'importe quelle* valeur numérique (non négative) pour le test de Wald des contraintes. Bien que dans la plupart des contextes économétriques il existe des paramétrisations qui apparaissent être plus *naturelles* que d'autres, et bien qu'on puisse espérer que l'utilisation de ces paramétrisations naturelles mèneraient à une inférence plus sûre que l'utilisation d'inférences moins naturelles, il existe une petite évidence que cet espoir est bien plus désireux immédiatement que l'on quitte le contexte des modèles de régression linéaires.

Examinons maintenant le manque d'invariance des statistiques de Wald un peu plus étroitement. Cela peut être vu à partir des expressions (13.03), (13.05), et (13.06) pour les trois statistiques classiques que c'est seulement dans la statistique de Wald  $W$  qu'apparaît la forme explicite des contraintes, à travers  $\hat{\mathbf{r}}$  et  $\hat{\mathbf{R}}$ . Si nous supposons, comme d'habitude, que les deux vecteurs paramétriques  $\boldsymbol{\theta}$  et  $\boldsymbol{\phi}$  correspondent au même DGP, tel que  $\boldsymbol{\phi} = \boldsymbol{\eta}(\boldsymbol{\theta})$ , alors les contraintes  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$  peuvent être exprimées dans des termes  $\boldsymbol{\phi}$  par la formule  $\mathbf{r}'(\boldsymbol{\phi}) = \mathbf{0}$ , où

$$\mathbf{r}'(\boldsymbol{\phi}) = \mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}(\boldsymbol{\eta}^{-1}(\boldsymbol{\phi})). \quad (13.75)$$

Ainsi,  $\mathbf{r}'$  peut être représenté comme la décomposition de deux applications  $\mathbf{r}$  et  $\boldsymbol{\eta}^{-1}$ :

$$\mathbf{r}' = \mathbf{r} \circ \boldsymbol{\eta}^{-1},$$

et (13.75) peut être écrit de manière équivalente comme

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{r}'(\boldsymbol{\eta}(\boldsymbol{\theta})). \quad (13.76)$$

La matrice  $\mathbf{R}'(\boldsymbol{\phi})$  est la matrice Jacobienne associée à l'application  $\mathbf{r}'$ , et aussi par la différenciation (13.76) par rapport à  $\boldsymbol{\theta}$  nous obtenons

$$\mathbf{R}(\boldsymbol{\theta}) \equiv D_{\boldsymbol{\theta}}\mathbf{r}(\boldsymbol{\theta}) = D_{\boldsymbol{\theta}}\mathbf{r}'(\boldsymbol{\eta}(\boldsymbol{\theta})) = D_{\boldsymbol{\phi}}\mathbf{r}'(\boldsymbol{\phi})D_{\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathbf{R}'(\boldsymbol{\phi})\mathbf{J}(\boldsymbol{\theta}), \quad (13.77)$$

par la règle ?en chaine? et (13.64). L'utilisation de (13.76), (13.77), et (13.69) dans l'expression (13.05) pour  $W$  donne alors la statistique pour la paramétrisation  $\boldsymbol{\phi}$ :

$$\begin{aligned} W' &\equiv n(\hat{\mathbf{r}}')^{\top}(\hat{\mathbf{R}}'(\hat{\mathbf{J}}')^{-1}(\hat{\mathbf{R}}')^{\top})^{-1}\hat{\mathbf{r}}' \\ &= n\hat{\mathbf{r}}^{\top}(\hat{\mathbf{R}}\hat{\mathbf{J}}^{-1}\hat{\mathbf{J}}\hat{\mathbf{J}}^{-1}\hat{\mathbf{J}}^{\top}(\hat{\mathbf{J}}^{\top})^{-1}\hat{\mathbf{R}}^{\top})^{-1}\hat{\mathbf{r}} \\ &= n\hat{\mathbf{r}}^{\top}(\hat{\mathbf{R}}\hat{\mathbf{J}}^{-1}\hat{\mathbf{R}}^{\top})^{-1}\hat{\mathbf{r}} = W, \end{aligned} \quad (13.78)$$

où les quantités ? primed? sont évaluées en  $\hat{\boldsymbol{\phi}}$  et celles ?unprimed? en  $\hat{\boldsymbol{\theta}}$ .

Ainsi où est le problème? La statistique  $W$  semble invariante selon son mode de calcul! La difficulté est que, précisément sur le compte de l'apparence explicite de  $\mathbf{r}$  et  $\mathbf{R}$  dans  $W$ , il est possible de reparamétriser, non seulement les paramètres du modèle,  $\boldsymbol{\theta}$  mais aussi les réels contraintes. Supposons que nous imaginons changer le paramètre de l'exemple de la dernière section de  $\sigma^2$  à  $\sigma$ . Si la contrainte  $\sigma^2 = 1$  est reformulée selon (13.76), elle devient  $\sigma^2 = (\sqrt{\sigma^2})^2 = 1$ . De manière similaire, si nous écrivons  $\tau = \log \sigma$ , la contrainte est  $e^{2\tau} = 1$ . Si nous avons utilisé la contrainte dans d'autres de ces formes, alors  $W$  serait en effet invariante sous les reparamétrisations, par (13.78). Mais ceci n'est pas serait probable de faire. Habituellement la contrainte serait écrite soit comme  $\sigma = 1$  soit comme  $\tau = 0$ . Alors, comme nous allons le montrer maintenant, la statistique n'est plus invariante.

Les  $r$  contraintes  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$  peuvent être exprimées de nombreuses manières différentes. Si  $\mathbf{p}$  est une quelconque application de  $\mathbb{R}^r$  to  $\mathbb{R}^r$  qui s'applique à l'origine et seulement à l'origine dans l'origine, alors, pour n'importe  $\mathbf{x} \in \mathbb{R}^r$ ,  $\mathbf{p}(\mathbf{x}) = \mathbf{0}$  si et seulement si  $\mathbf{x} = \mathbf{0}$ . Ainsi, les contraintes  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$  sont complètement équivalentes aux contraintes  $\mathbf{p}(\mathbf{r}(\boldsymbol{\theta})) = \mathbf{0}$ . Si nous écrivons  $\mathbf{q}$  pour la composition  $\mathbf{p} \circ \mathbf{r}$ , alors  $\mathbf{q}$  s'applique  $\mathbb{R}^k$  dans  $\mathbb{R}^r$ , et exactement la même sous série de l'espace paramétrique  $\Theta$  est définie par l'imposition  $\mathbf{q}(\boldsymbol{\theta}) = \mathbf{0}$  comme par l'imposition  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ . Dans ce sens, nous pouvons appeler les contraintes  $\mathbf{q}(\boldsymbol{\theta}) = \mathbf{0}$  une reparamétrisation de  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ .

Afin de formuler une statistique de Wald statistic pour ces contraintes de reparamétrisations, nous avons besoin de la matrice Jacobienne de  $\mathbf{q}$ , que nous appellerons  $\mathbf{Q}$ . C'est

$$\mathbf{Q}(\boldsymbol{\theta}) \equiv D_{\boldsymbol{\theta}}\mathbf{q}(\boldsymbol{\theta}) = D_{\boldsymbol{\theta}}(\mathbf{p}(\mathbf{r}(\boldsymbol{\theta}))) = D_{\mathbf{r}}\mathbf{p}(\mathbf{r}(\boldsymbol{\theta})) D_{\boldsymbol{\theta}}\mathbf{r}(\boldsymbol{\theta}) = D_{\mathbf{r}}\mathbf{p}(\mathbf{r}(\boldsymbol{\theta}))\mathbf{R}(\boldsymbol{\theta}).$$

Par conséquent, la statistique de Wald, en notation évidemment, est

$$\begin{aligned} W'' &\equiv n\hat{\mathbf{q}}^{\top}(\hat{\mathbf{Q}}\hat{\mathbf{J}}^{-1}\hat{\mathbf{Q}}^{\top})^{-1}\hat{\mathbf{q}} \\ &= n\mathbf{p}^{\top}(\hat{\mathbf{r}})(D_{\mathbf{r}}\mathbf{p}(\hat{\mathbf{r}})\hat{\mathbf{R}}\hat{\mathbf{J}}^{-1}\hat{\mathbf{R}}^{\top}D_{\mathbf{r}}^{\top}\mathbf{p}(\hat{\mathbf{r}}))^{-1}\mathbf{p}(\hat{\mathbf{r}}), \end{aligned} \quad (13.79)$$

qui n'est pas en général égale à la statistique d'origine  $W$  ou à la statistique  $W' = W$  que nous avons obtenue dans (13.78) lorsque nous avons reparamétrisé le modèle mais pas les contraintes.

Il existe encore un cas important pour lequel  $W''$  dans (13.79) est égal à  $W$ , à savoir, le cas d'une application linéaire non singulière  $\mathbf{p}$ . (Une telle application s'applique automatiquement à l'origine et seulement à l'origine, naturellement.) Si  $\mathbf{p}$  est linéaire, alors il n'y a pas de réelle différence entre  $\mathbf{p}$  elle-même et sa Jacobienne  $D_{\mathbf{r}}\mathbf{p}$ . Nous pouvons alors écrire pour n'importe quel  $\boldsymbol{\theta} \in \Theta$  que

$$\mathbf{p}(\mathbf{r}(\boldsymbol{\theta})) = D_{\mathbf{r}}\mathbf{p}(\mathbf{r}(\boldsymbol{\theta}))\mathbf{r}(\boldsymbol{\theta}). \quad (13.80)$$

Ceci doit être interprété pour dire que le vecteur  $\mathbf{r} - \mathbf{p}(\mathbf{r}(\boldsymbol{\theta}))$  est égal au produit de la matrice  $D_{\mathbf{r}}\mathbf{p}(\mathbf{r}(\boldsymbol{\theta}))$  de dimension  $r \times r$  et le vecteur  $\mathbf{r}(\boldsymbol{\theta})$  de dimension  $r$ . L'utilisation de (13.80) dans (13.79) fait coïncider  $W''$  avec  $W$  et  $W'$ .

Avant de conclure dans cette section, nous devrions noter une propriété d'invariance complémentaire, mais d'une sorte plutôt différente de celles que nous avons étudiées jusqu'ici. Cette propriété, qui a été révélée par Godfrey (1981), est très particulière pour le test LM; il n'existe rien d'analogue pour les tests LR et Wald. Il en ressort que, lorsqu'une hypothèse nulle donnée est soumise à un test, la même statistique LM peut être obtenue pour deux ou trois hypothèses alternatives différentes si ces dernières sont **localement équivalentes**. Nous avons déjà rencontré un exemple similaire de ce phénomène au Chapitre 10, dans lequel nous avons vu qu'une et même statistique de test est générée lorsqu'un modèle de régression est testé, au moyen

d'une GNR, pour la présence d'erreurs soient  $AR(p)$  soient  $MA(p)$ . Une implication importante de l'équivalence locale est la suivante. Si deux hypothèses alternatives sont localement équivalentes, alors pour n'importe quel mise en marche de DGP qui appartient à une autre alternative, n'importe quel test asymptotique pour lequel l'alternative explicite est une d'entre-elles aura une ARE relative à l'unité pour n'importe quel test pour lequel l'alternative explicite est l'autre.

Nous examinons maintenant juste quel aspect des différentes hypothèses alternatives est responsable de cette invariance de la statistique LM. Rappelons de (13.03) qu'une statistique LM est constituées de deux éléments, à savoir, le gradient de la fonction logvraisemblance et la matrice d'information, évalués tous deux en des estimations contraintes ML. Ces estimations dépendent seulement de l'hypothèse nulle sous test et sont alors invariantes au changements dans les hypothèses alternatives. Plus loin, la matrice d'information est définies comme l'espérance du produit extérieur du gradient avec lui-même; consulter (13.67). Ainsi si, pour un échantillon donné, nous testons la même hypothèse nulle contre deux alternatives différentes, et le gradient tend à être le même pour les deux alternatives, alors la statistique LM entière sera la même. Ce résultat suppose que nous utilisons la forme score efficace du test LM. Si nous avons basé le test sur des estimations de la matrice d'information, les deux statistiques LM ne pourraient pas être numériquement les mêmes, bien qu'elles le seraient asymptotiquement.

Géométriquement, deux hypothèses alternatives différentes sont localement équivalentes si elles **se touchent** à l'hypothèse nulle. Par ceci, nous ne signifions pas que les deux hypothèses alternatives donnent les mêmes valeurs de leurs fonctions de logvraisemblance respectives lorsqu'elles sont contraintes par l'hypothèse nulle, comme cela sera toujours le cas, mais aussi que les gradients des deux fonctions de logvraisemblance sont les mêmes, car les gradients sont *tangents* pour les deux modèles qui se touchent au modèle nul. Dans ces circonstances, les deux tests LM doivent être numériquement identiques.

Que signifie pour les deux modèles se toucher, ou, utiliser le terme non géométrique pour la propriété, être localement équivalent? Une définition circulaire serait simplement que leurs gradients soient les mêmes en tous les DGP que les deux modèles interceptent. Statistiquement, cela signifie que si on part seulement légèrement de l'hypothèse nulle tout en respectant une des deux hypothèses alternatives, alors on part d'une autre hypothèse alternative par un montant qui est du second ordre de petites quantités. Par exemple, un processus  $AR(1)$  caractérisé par un petit paramètre autorégressif  $\rho$  diffère d'un certain processus  $MA(1)$  à une étendue seulement proportionnelle à  $\rho^2$ . To Pour prouver ceci formellement, une définition formelle de la distance comprise entre les deux DGP devrait se substituer, mais notre définition circulaire antérieur est une définition opérationnelle: si le gradient  $\tilde{g}^1$  calculé pour la première alternative est le même que le gradient  $\tilde{g}^2$  pour la seconde, alors les deux alternatives se touchent à l'hypothèse nulle. Il serait clair maintenant

que cette exigence soit très forte: Elle est suffisante si les éléments de  $\tilde{\mathbf{g}}^2$  sont tous des combinaisons linéaires de celles de  $\tilde{\mathbf{g}}^1$  et vice versa. Un exemple de cette dernière possibilité est fourni par l'équivalence locale, autour de ? l'hypothèse? nulle des erreurs à bruit blanc, des modèles de régression avec des erreurs ARMA( $p, q$ ) d'un côté et avec des erreurs AR( $p+q$ ) de l'autre; voir la Section 10.8. Pour plus d'exemples, consulter Godfrey (1981) et Godfrey et Wickens (1982).

Les aspects à la fois algébriques et géométriques de l'invariance des tests LM sous équivalence locale sont exprimés au moyen d'une simple remarque: le test LM peut être construit seulement sur la base des estimations contraintes ML et des dérivées *premières* de la fonction vraisemblance évaluées en ces estimations. Ceci implique que le test LM ne tient pas compte de la courbure de l'hypothèse alternative proche de l'hypothèse nulle.

Nous pouvons résumer les résultats de cette comme suit:

1. Le test LR dépend seulement de deux fonctions de logvraisemblance maximisées. Il ne peut pas peut alors dépendre soit du modèle de reparamétrisation soit de la manière dont les contraintes sont formulées. Il en terme de ces paramètres.
2. La forme score efficace de la statistique LM est construite de deux éléments, le gradient et la matrice d'information, qui se modifient sous la reparamétrisation, mais d'une telle manière que la statistique de test elle-même est invariante, non seulement sous des reparamétrisations mais aussi sous différents choix, localement équivalents, de l'hypothèse alternative. Si la matrice d'information elle-même doit être estimée, alors la dépendance de reparamétrisation peut apparaître, comme nous l'avons vu quand la matrice d'information a été estimée en utilisant le Hessien et considérée une reparamétrisation non linéaire. Cependant, l'estimateur de la matrice d'information OPG de (8.48) transforme aussi sous des reparamétrisations de telle manière qu'il laisse la statistique LM invariante; il s'agit d'un bon exercice que de montrer ceci.
3. Le test de Wald peut être dépendant à la paramétrisation pour la même raison comme pour le test LM mais peut en plus l'être pour une raison différente, pas directement à partir des paramètres du modèle mais à travers la manière dans laquelle ils sont utilisés pour formuler les contraintes.
4. Si une reparamétrisation ou reformulation des contraintes est linéaire, cela n'affecte pas la valeur de n'importe laquelle des statistiques de test classiques.



### 13.7 LA RÉGRESSION DU PRODUIT EXTÉRIEUR DU GRADIENT

Nous avons remarqué dans l'introduction de ce chapitre que la régression Gauss-Newton n'est pas généralement applicable aux modèles estimés par maximum de vraisemblance. En vue de l'inutilité extrême de la GNR pour le calcul des statistiques de test dans le contexte des modèles de régression non linéaire, il est d'un grand intérêt de voir si les autres régressions artificielles comportant des propriétés similaires sont convenables dans le contexte des modèles estimés par maximum de vraisemblance.

Une remarque préliminaire et évidente: aucune régression, artificielle ou autre, est nécessaire pour mettre en exécution le test LR. Car n'importe quel programme capable de produire des estimations ML donnera certainement aussi la fonction de logvraisemblance maximisée, il peut n'y avoir aucun obstacle à opérer un test LR à moins qu'il ait une quelconque difficulté dans l'estimation soit du modèle contraint, soit non contraint. Dans certains cas, il n'y a pas de telle difficulté, et alors le test LR est presque toujours la procédure choisie. Cependant, il existe certaines occasions où un des deux modèles est plus facile à estimer que l'autre, et où on espérerait utiliser soit le test LM soit le test de Wald pour éviter certaines difficultés d'estimation. Une autre possibilité est que l'hypothèse alternative peut être implicite plutôt que d'être associée à un modèle paramétré bien défini qui inclue l'hypothèse nulle comme un cas spécial. Nous avons vu dans le contexte de la GNR que beaucoup de tests diagnostiques tombe dans cette catégorie. Lorsque l'hypothèse alternative est implicite, on espérerait presque toujours utiliser un test LM.

Dans le contexte de régression, la GNR fournit un moyen de calculer des statistiques de test basées sur le principe de LM. A vrai dire, comme nous l'avons vu à la Section 6.7, elle peut être utilisée pour calculer des statistiques de test classiques basées sur n'importe quelles estimations convergentes au taux  $n^{1/2}$ . Nous introduisons maintenant une nouvelle régression artificielle, appelée la **régression du produit-extérieur-au-gradient**, ou la **régression OPG** pour faire court, qui peut être utilisée pour n'importe quel modèle estimé par maximum de vraisemblance. la régression OPG a été utilisée premièrement comme un moyen de calcul des statistiques de test par Godfrey and Wickens (1981). Cette régression artificielle, qui est très facile à établir pour la plupart des modèles estimés par maximum de vraisemblance, peut être utilisée dans la mêmes cas que la GNR: vérification des conditions du premier ordre pour la fonction de logvraisemblance, l'estimation de la matrice de covariance, l'estimation efficace en une étape, du plus grand intérêt immédiat pour le calcul des statistiques de test.

Supposons que nous sommes intéressés au modèle paramétré (13.01). Soit  $\mathbf{G}(\boldsymbol{\theta})$  la matrice CG associée à la fonction de logvraisemblance  $\ell^n(\boldsymbol{\theta})$ , avec comme élément type

$$G_{ti}(\boldsymbol{\theta}) \equiv \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i}; \quad t = 1, \dots, n, \quad i = 1, \dots, k,$$

où  $k$  est le nombre d'éléments dans le vecteur paramétrique  $\boldsymbol{\theta}$ . Alors la régression OPG associée au modèle (13.01) peut être écrite comme

$$\boldsymbol{\iota} = \mathbf{G}(\boldsymbol{\theta})\mathbf{c} + \text{résidus}. \quad (13.81)$$

Ici,  $\boldsymbol{\iota}$  est un vecteur de dimension  $n$ —duquel chaque élément est unitaire et  $\mathbf{c}$  est un vecteur de dimension  $k$ —des paramètres artificiels. Le produit de la matrice des régresseurs avec la régressande est le gradient  $\mathbf{g}(\boldsymbol{\theta}) \equiv \mathbf{G}^\top(\boldsymbol{\theta})\boldsymbol{\iota}$ . La matrice des sommes des carrés et les produits croisés des régresseurs,  $\mathbf{G}^\top(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})$ , lorsqu'elle est divisée par  $n$ , estime de manière convergente la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$ . Ces deux caractéristiques sont essentiellement tout ce qu'il faut pour que (13.81) soit une régression artificielle valide.<sup>6</sup> comme avec la GNR, les régresseurs de la régression OPG dépendent du vecteur  $\boldsymbol{\theta}$ . Par conséquent, avant que la régression artificielle ne tourne, ces régresseurs doivent être évalués en certain vecteur paramétrique choisi.

Un choix possible pour ce vecteur paramétrique est  $\hat{\boldsymbol{\theta}}$ , l'estimateur ML pour le modèle (13.01). Dans ce cas, la matrice du régresseur est  $\hat{\mathbf{G}} \equiv \mathbf{G}(\hat{\boldsymbol{\theta}})$  et les estimations paramétriques artificielles, qui sont désignées par  $\hat{\mathbf{c}}$ , sont identiquement zéro:

$$\hat{\mathbf{c}} = (\hat{\mathbf{G}}^\top \hat{\mathbf{G}})^{-1} \hat{\mathbf{G}}^\top \boldsymbol{\iota} = (\hat{\mathbf{G}}^\top \hat{\mathbf{G}})^{-1} \hat{\mathbf{g}} = \mathbf{0}.$$

Comme ici  $\hat{\mathbf{g}}$  est la fonction de logvraisemblance du gradient évaluée en  $\hat{\boldsymbol{\theta}}$ , la dernière égalité du de-dessus est une conséquence des conditions du premier ordre pour le maximum de vraisemblance. Comme avec la GNR, alors, la mise en marche de la régression OPG avec  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$  fournit une simple manière de tester comment, en fait, sont bien satisfaites les conditions du premier ordre par une série d'estimations calculées au moyen d'un certain programme informatique. Les statistiques  $t$  fournissent encore la vérification la plus convenable. Elles ne devraient pas excéder un nombre aux alentours de  $10^{-2}$  ou  $10^{-3}$  en valeur absolue si une bonne approximation pour le maximum a été trouvée.

Comme les estimations  $\hat{\mathbf{c}}$  pour la régression (13.81) sont zéro lorsque les régresseurs sont  $\hat{\mathbf{G}}$ , ces régresseurs n'ont aucun pouvoir explicatif pour  $\boldsymbol{\iota}$ , et la somme des résidus au carré est alors égale au total de la somme des carrés. Parce que cette dernière est

$$\boldsymbol{\iota}^\top \boldsymbol{\iota} = \sum_{t=1}^n 1 = n,$$

l'estimation ML de la variance des résidus dans (13.81) est juste l'unité:

$$\frac{1}{n} \text{SSR} = \frac{1}{n} \boldsymbol{\iota}^\top \boldsymbol{\iota} = \frac{1}{n} n = 1.$$

<sup>6</sup> Des conditions précises pour une régression pour être appelée "artificiel" sont fournies par Davidson and MacKinnon (1990); consulter la Section 14.4.

L'estimation de la variance OLS, qui est  $SSR/(n-k) = n/(n-k)$ , est asymptotiquement équivalente à ceci, mais l'exposé sera simplifié si nous supposons que l'estimation ML est utilisée. L'estimation de la matrice de covariance pour le vecteur  $\hat{c}$  de (13.81) est alors

$$(\hat{G}^\top \hat{G})^{-1}.$$

C'est cette expression qui donne à la régression OPG son nom, pour son inverse, est précisément l'estimateur OPG de la matrice d'information; consulter (8.48) et (8.50).<sup>7</sup> Il s'en suit que, comme avec la GNR,  $n^{-1}$  fois l'estimateur de la matrice de covariance de la régression OPG est asymptotiquement égal à la matrice de covariance de  $n^{1/2}(\hat{\theta} - \theta_0)$ .

La propriété juste établie n'est la seule partagée par les régressions Gauss-Newton et OPG. Nous établissons maintenant deux propriétés complémentaires de la régression OPG qui sont en fait partagées par toutes les régressions auxquelles nous attribuons le nom d'"artificielles." La première de ces propriétés est qu'elle nous permet d'utiliser des régressions artificielles pour exécuter l'estimation efficace en une étape. selon cette propriété, si la régression OPG (13.81) est évaluée en un certain vecteur paramétrique  $\hat{\theta}$  qui est convergent au taux  $n^{1/2}$  pour  $\theta_0$ , tel que  $\hat{\theta} - \theta_0 = O(n^{-1/2})$ , alors les estimations paramétriques artificielles  $\hat{c}$  sont telles que

$$n^{1/2}\hat{c} \stackrel{a}{=} n^{1/2}(\hat{\theta} - \hat{\theta}), \quad (13.82)$$

où  $\hat{\theta}$  est l'estimateur ML de  $\theta$ . Ce résultat est essentiellement identique à celui démontré dans le cadre de la GNR dans la section 6.6.

Le résultat (13.82) est important. Grâce à lui, nous pouvons passer en une étape à partir de n'importe quel estimateur convergent au taux  $n^{1/2}$   $\hat{\theta}$  d'un estimateur équivalent asymptotiquement à un estimateur efficace asymptotiquement  $\hat{\theta}$ . L'estimateur en une étape  $\hat{\theta}$  défini par  $\hat{\theta} \equiv \hat{\theta} + \hat{c}$  comporte la propriété que

$$n^{1/2}(\hat{\theta} - \theta_0) = n^{1/2}(\hat{\theta} - \theta_0) + o(1), \quad (13.83)$$

comme cela peut être vu directement de (13.82). Comme l'équivalence asymptotique de  $\hat{\theta}$  et  $\hat{\theta}$  demande des facteurs de  $n^{1/2}$  qui apparaissent dans (13.83), il peut être vu pourquoi nous espérons prouver (13.82) avec un facteur de  $n^{1/2}$  de chaque côté de l'équation, plutôt que le résultat équivalent en apparence que  $\hat{c} \stackrel{a}{=} \hat{\theta} - \hat{\theta}$ . Bien que ce résultat soit certainement vrai, plus faible que (13.82), parce qu'il implique simplement que  $\hat{\theta} - \hat{\theta} = o(1)$ , tandis que (13.82) implique que  $\hat{\theta} - \hat{\theta} = o(n^{-1/2})$ .

<sup>7</sup> As nous avons noté à la Section 8.6, que certains auteurs se réfèrent à l'estimateur OPG de la matrice d'information comme l'estimateur BHHH estimator, après Berndt, Hall, Hall, et Hausman (1974), qui ont défendu son utilisation, bien qu'ils n'ont rendu explicite l'utilisation de la régression OPG elle-même.

La preuve de (13.82) est à la fois simple et lumineuse. Un développement de Taylor du gradient  $\dot{\mathbf{g}} \equiv \mathbf{g}(\hat{\boldsymbol{\theta}})$  autour de  $\boldsymbol{\theta}_0$  donne

$$n^{-1/2}\dot{\mathbf{g}} = n^{-1/2}\mathbf{g}_0 + n^{-1}\mathbf{H}(\boldsymbol{\theta}_0)n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O(n^{-1/2}),$$

où, comme d'habitude,  $\mathbf{H}(\boldsymbol{\theta})$  désigne le Hessian de la fonction de logvraisemblance  $\ell(\boldsymbol{\theta})$ . Si maintenant nous développons  $\hat{\mathbf{g}}$ , qui est zéro par les conditions du premier ordre pour un maximum de la vraisemblance en  $\hat{\boldsymbol{\theta}}$ , nous obtenons

$$\mathbf{0} = n^{-1/2}\mathbf{g}_0 + n^{-1}\mathbf{H}(\boldsymbol{\theta}_0)n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O(n^{-1/2}).$$

En soustrayant la dernière des deux équations et en notant que  $\dot{\mathbf{g}} = \dot{\mathbf{G}}^\top \boldsymbol{\iota}$ , nous trouvons que

$$n^{-1/2}\dot{\mathbf{G}}^\top \boldsymbol{\iota} = n^{-1}\mathbf{H}(\boldsymbol{\theta}_0)n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O(n^{-1/2}). \quad (13.84)$$

Par l'égalité de la matrice d'information,  $n^{-1}\mathbf{H}(\boldsymbol{\theta}_0) = -\mathcal{J}_0 + o(1)$ . Comme, par la convergence de  $\hat{\boldsymbol{\theta}}$ , nous avons  $n^{-1}\dot{\mathbf{G}}^\top \dot{\mathbf{G}} = \mathcal{J}_0 + o(1)$ , nous pouvons remplacer  $n^{-1}\mathbf{H}(\boldsymbol{\theta}_0)$  dans (13.84) par  $-n^{-1}\dot{\mathbf{G}}^\top \dot{\mathbf{G}}$  pour obtenir

$$n^{-1/2}\dot{\mathbf{G}}^\top \boldsymbol{\iota} = (n^{-1}\dot{\mathbf{G}}^\top \dot{\mathbf{G}})n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o(1).$$

Le résultat (13.82) provient maintenant directement de la prémultiplication par  $(n^{-1}\dot{\mathbf{G}}^\top \dot{\mathbf{G}})^{-1}$ .

Une seconde propriété des régressions artificielles est celle qui permet leur utilisation dans le calcul des statistiques LM. Lorsqu'une régression artificielle qui satisfait cette propriété est évaluée en  $\hat{\boldsymbol{\theta}}$  à un taux de convergence  $n^{1/2}$ ,  $n$  fois le  $R^2$  non centré calculé à partir d'elle, est asymptotiquement équivalent à

$$\frac{1}{n}\dot{\mathbf{g}}^\top \mathcal{J}_0^{-1}\dot{\mathbf{g}}.$$

Ce résultat est très facile à prouver pour la régression OPG. Le  $R^2$  est le ratio de la somme expliquée des carrés (ESS) à la somme totale des carrés (TSS), et ainsi  $nR^2$  est le ratio ESS/(TSS/ $n$ ). Nous avons vu que TSS/ $n$  était égale à 1. Ceci signifie que  $nR^2$  est juste la somme expliquée des carrés:

$$nR^2 = \boldsymbol{\iota}^\top \dot{\mathbf{G}}(\dot{\mathbf{G}}^\top \dot{\mathbf{G}})^{-1}\dot{\mathbf{G}}^\top \boldsymbol{\iota} = \frac{1}{n}\dot{\mathbf{g}}^\top (n^{-1}\dot{\mathbf{G}}^\top \dot{\mathbf{G}})^{-1}\dot{\mathbf{g}}. \quad (13.85)$$

Ceci complète la preuve, car  $n^{-1}\dot{\mathbf{G}}^\top \dot{\mathbf{G}} \rightarrow \mathcal{J}_0$ .

En utilisant ce résultat, nous voyons que la statistique LM (13.03) peut être calculée très facilement. Certains programmes de régression ne mentionnent pas le  $R^2$  non centré, et certains ne mentionnent même pas la somme des carrés expliqués. Par conséquent, les deux moyens les plus naturels de calculer la statistique de test peuvent ne pas être convenables. Comme tous les programmes mentionnent la somme des résidus au carré, une troisième manière est d'utiliser le fait que, pour la régression OPG,

$$nR^2 = \text{ESS} = n - \text{SSR}.$$

pour calculer la statistique LM, alors, on peut simplement calculer  $n$  moins la somme des résidus au carré pour la régression OPG évaluée en  $\tilde{\theta}$ . Bien que ceci soit la manière la plus simple à partir de laquelle une statistique basée sur le principe LM puisse être calculée au moyen d'une régression OPG, elle n'est pas la seule. Par exemple, s'il n'existe qu'une contrainte et que  $\tilde{G}$  puisse être partitionné tel que  $k-1$  colonnes soient orthogonales à  $\iota$  et qu'une colonne ne lui soit pas orthogonal, on peut utiliser un test ordinaire  $t$  sur le coefficient de ce dernier. Les détails, suivent très étroitement ceux de la GNR, sont laissés en exercice pour le lecteur.

Il est aussi possible de d'utiliser une régression OPG, ou de fait n'importe quelle régression artificielle qui satisfassent les propriétés du-dessus, pour calculer les **statistiques de test**  $C(\alpha)$ , basées sur n'importe quelles estimations convergentes au taux  $n^{1/2}$  qui satisfont l'hypothèse nulle. Le test  $C(\alpha)$ , que nous avons mentionné aux Chapitres 6, 7, et 11, a été proposé en premier par Neyman (1959); consulter Neyman et Scott (1966), Moran (1970), Breusch et Pagan (1980), Smith (1987), et Dagenais et Dufour (1991) pour des discussions plus détaillées et des applications. Le test  $C(\alpha)$  peut être regarder comme un test classique. Bien qu'il soit moins bien connu que les tests LM, LR, et Wald, il leur est, comme nous le démontrons maintenant, asymptotiquement équivalent.

Supposons que  $\theta$  soit partitionnée comme  $[\theta_1 : \theta_2]$ , que le vecteur gradient  $g(\theta)$  et la matrice d'information  $J(\theta)$  soient partitionnés de la même manière, et que les contraintes sur  $\theta$  peuvent être écrites comme  $\theta_2 = \theta_2^0$ . Alors les statistiques de test  $C(\alpha)$  peuvent être écrites de différentes manière, parmi lesquelles la plus simple est

$$C(\alpha) \equiv \frac{1}{n} \tilde{g}^\top \tilde{J}^{-1} \tilde{g} - \frac{1}{n} \tilde{g}_1^\top (\tilde{J}_{11})^{-1} \tilde{g}_1, \quad (13.86)$$

où toutes les quantités sont évaluées aux estimations  $\hat{\theta} = [\hat{\theta}_1 : \hat{\theta}_2^0]$  convergentes au taux  $n^{1/2}$  qui satisfont l'hypothèse nulle.

Il est facile de voir que le test  $C(\alpha)$  est asymptotiquement équivalent aux autres tests classiques. Les approximations des séries de Taylor au premier ordre de  $n^{-1/2}g(\hat{\theta})$  et  $n^{-1/2}g_1(\hat{\theta})$  autour de  $\tilde{\theta}$ , combinées à l'égalité de la matrice d'information, donnent les résultats

$$n^{-1/2}g(\hat{\theta}) \stackrel{a}{\approx} n^{-1/2}g(\tilde{\theta}) - J(\theta_0)n^{1/2}(\hat{\theta} - \tilde{\theta}) \quad \text{and} \quad (13.87)$$

$$n^{-1/2}g_1(\hat{\theta}) \stackrel{a}{\approx} -J_{11}(\theta_0)n^{1/2}(\hat{\theta}_1 - \tilde{\theta}_1). \quad (13.88)$$

En dérivant (13.88), nous avons utilisé les conditions du premier ordre pour  $\tilde{\theta}_1$  et le fait que  $\theta_2 = \tilde{\theta}_2$ . En utilisant (13.87), le premier terme dans (13.86) est

$$\begin{aligned} \frac{1}{n} \dot{\mathbf{g}}^\top \dot{\mathbf{J}}^{-1} \dot{\mathbf{g}} &\stackrel{a}{=} (n^{-1/2} \tilde{\mathbf{g}} - \mathcal{J} n^{1/2} (\dot{\theta} - \tilde{\theta}))^\top \mathcal{J}^{-1} (n^{-1/2} \tilde{\mathbf{g}} - \mathcal{J} n^{1/2} (\dot{\theta} - \tilde{\theta})) \\ &= \frac{1}{n} \tilde{\mathbf{g}}^\top \mathcal{J}^{-1} \tilde{\mathbf{g}} + n (\dot{\theta}_1 - \tilde{\theta}_1)^\top \mathcal{J}_{11} (\dot{\theta}_1 - \tilde{\theta}_1). \end{aligned} \quad (13.89)$$

Il y a seulement deux termes dans la seconde ligne de (13.89), parce que les deux autres termes concernent les produits internes de  $\tilde{\mathbf{g}}$  avec  $\dot{\theta} - \tilde{\theta}$ . Comme  $\tilde{\mathbf{g}}_1 = \mathbf{0}$ , ces produits internes sont zéro. En utilisant (13.88) et la seconde ligne de (13.89), nous obtenons

$$\frac{1}{n} \dot{\mathbf{g}}^\top \dot{\mathbf{J}}^{-1} \dot{\mathbf{g}} \stackrel{a}{=} \frac{1}{n} \tilde{\mathbf{g}}^\top \mathcal{J}^{-1} \tilde{\mathbf{g}} + \frac{1}{n} \dot{\mathbf{g}}_1^\top (\mathcal{J}_{11})^{-1} \dot{\mathbf{g}}_1. \quad (13.90)$$

Comme le second terme dans (13.86) est moins une estimation convergente du second terme dans (13.90), il s'en suit que

$$C(\alpha) \equiv \frac{1}{n} \dot{\mathbf{g}}^\top \dot{\mathbf{J}}^{-1} \dot{\mathbf{g}} - \frac{1}{n} \dot{\mathbf{g}}_1^\top (\mathcal{J}_{11})^{-1} \dot{\mathbf{g}}_1 \stackrel{a}{=} \frac{1}{n} \tilde{\mathbf{g}}^\top \mathcal{J}^{-1} \tilde{\mathbf{g}}.$$

Ainsi, nous concluons que la statistique  $C(\alpha)$  est asymptotiquement équivalente à la statistique LM et désormais de toutes les statistiques de test classiques.

La statistique de test classique (13.86) est la différence entre deux formes quadratiques. En fait, cela ressemble à la différence entre deux statistiques LM. La première de celles-ci est asymptotiquement égale à  $nR^2$  à partir de la régression artificielle

$$\boldsymbol{\iota} = \dot{\mathbf{G}}_1 \mathbf{c}_1 + \dot{\mathbf{G}}_2 \mathbf{c}_2 + \text{résidus}, \quad (13.91)$$

et la seconde est asymptotiquement égale à  $nR^2$  à partir de la régression artificielle

$$\boldsymbol{\iota} = \dot{\mathbf{G}}_1 \mathbf{c}_1 + \text{résidus}. \quad (13.92)$$

Le  $R^2$  de cette régression serait zéro si  $\dot{\theta} = \tilde{\theta}$ , par les conditions du premier ordre pour  $\tilde{\theta}$ , mais ne sera pas zéro généralement pour n'importe quel autre choix de  $\dot{\theta}$ .

La statistique de test (13.86) est asymptotiquement égale à  $n$  fois la différence entre le  $R^2$  de (13.91) et de (13.92). Ainsi, nous voyons que les tests LM basés sur la régression OPG sont juste des cas spéciaux des tests  $C(\alpha)$  basés sur cette régression. La différence est que, parce que  $nR^2$  de (13.92) n'est généralement pas zéro, on ne peut pas simplement utiliser  $nR^2$  de (13.91) comme la statistique de test dans le cas le plus général. L'intuition que souligne ce résultat est très simple. De (13.82), nous voyons que les estimations OLS de  $\mathbf{c}_2$  dans (13.91) sont asymptotiquement équivalente aux

estimations non contraintes ML  $\hat{\theta}_2$  moins  $\theta_2^0$ . Ainsi, il est à peine surprenant qu'un test pour  $c_2 = \mathbf{0}$  dans (13.91) devrait être équivalent à un test classique pour  $\theta_2 = \theta_2^0$ .

Comme nous l'avons noté à la Section 6.7, il est possible de calculer des statistiques basées sur principe de Wald en utilisant des régressions artificielles. Nous nous référerons à celles comme les **statistiques tout-comme-Wald**, parce qu'elles ne sont pas en général numériquement égales aux statistiques Wald calculées conventionnellement comme dans (13.05). Elles sont naturellement *asymptotiquement* égales à la statistique de test classique Wald et partagent avec elle la propriété d'être basées exclusivement sur les estimations ML du modèle non contraint. Malheureusement, elles partagent aussi la propriété d'être dépendant à la paramétrisation. Considérons la régression OPG correspondant au modèle non contraint évalué en  $[\hat{\theta}_1 : \theta_2^0]$ , un vecteur paramétrique qui, par construction, satisfait l'hypothèse nulle. Cette régression artificielle est

$$v = G_1(\hat{\theta}_1, \theta_2^0)c_1 + G_2(\hat{\theta}_1, \theta_2^0)c_2 + \text{résidus}. \quad (13.93)$$

Ceci est juste un cas spécial de la régression  $C(\alpha)$  (13.91), et n'importe quel test asymptotiquement valide de l'hypothèse artificielle  $c_2 = \mathbf{0}$  basée sur (13.93) fournit un test valide tout-comme-Wald.

Les tests LM,  $C(\alpha)$ , and tout-comme-Wald basés sur la régression OPG sont tellement simples que cela semble inviter à suggérer tous tests autres que le test LR peut être calculé le plus commodément au moyen d'une régression OPG. Cependant, comme il est clair de (13.85) pour le test LM, tous les tests basés sur la régression OPG utilisent l'estimateur des produit-extérieur-au-gradient de la matrice d'information. Bien que cet estimateur comporte un avantage d'être indépendant à la paramétrisation, de nombreuses expériences de Monte Carlo ont montré que ses propriétés d'échantillon fini sont presque toujours très différentes de ses propriétés asymptotiques nominales à moins que les tailles ne soient très grande, souvent de l'ordre de quelques milliers. En particulier, ces expériences suggèrent que les tests OPG ont souvent une taille loin par excès de leur taille asymptotique. Les véritables hypothèses nulles sont rejetées beaucoup trop souvent, dans certains mauvais cas spéciaux, presque à chaque fois. Consulter, parmi d'autres, Davidson and MacKinnon (1983a, 1985c, 1992a), Bera et McKenzie (1986), Godfrey, McAleer, et McKenzie (1988), et Chesher et Spady (1991). Bien que certaines expériences ont suggéré que les tests basés-OPG ont à ce sujet beaucoup plus de puissance que d'autres variantes des statistiques de test *si* une manière peut être trouvée pour corriger leur taille, personne n'a trouvé une quelconque manière facile et commode d'exécuter la nécessaire correction de taille.

Au vue de cette caractéristique plutôt navrante de la régression OPG, nous concluons cette section avec une ferme remontrance aux lecteurs qui l'utilisent avec grand soin. Dans la plupart des cas, il n'est pas risqué de conclure qu'une contrainte soit compatible avec les données si une statistique de test calculée en utilisant la régression OPG fait défaut pour rejeter l'hypothèse

nulle. Mais il est généralement risqué de conclure qu'une contrainte est incompatible avec les données si une statistique de test OPG rejette l'hypothèse nulle, du moins pas pour les échantillons de certaine taille ordinaire. Naturellement, si quelque chose est connu concernant les propriétés du test particulier OPG étant utilisé, peut-être comme un résultat des expériences de Monte Carlo, on peut alors être capable de d'établir des conclusions d'une statistique de test OPG qui rejette l'hypothèse nulle.

Cependant, la régression OPG serait importante même si personne ne l'a jamais utilisée pour calculer des statistiques de test. Son utilisation dans des calculs asymptotiques *théoriques* peut rendre de tels calculs beaucoup plus simples qu'ils ne pourraient l'être autrement. Deplus, comme nous le verrons dans les deux prochains chapitres il existe d'autres régressions artificielles, généralement pas tout à fait très applicables comme la régression OPG peut-être, et pas tout à fait très élémentaires à établir complètement, mais possédant généralement de meilleurs propriétés d'échantillon fini. Les résultats qui sont vrais pour la régression OPG sont aussi vrais pour ces autres régressions artificielles.

### 13.8 LECTURE COMPLÉMENTAIRE ET CONCLUSION

Les trois tests classiques, comme le mot "classiques" l'indique, possèdent une longue histoire et ont générés une formidable littérature; consulter Engle (1984) et Godfrey (1988) pour des références. Dans ce chapitre, nous avons essayer d'attirer l'attention sur des aspects communs des tests en soulignant la très foisonnante diversité des procédures de test de faire ressortir l'interprétation géométrique des tests. Une discussion plus simple de la géométrie des tests classiques peut être trouvée dans Buse (1982). Nous avons signalé qu'il existe une variable aléatoire commune vers laquelle toute les statistiques de test classiques tendent quand la taille d'échantillon tend vers l'infini et que la distribution de cette variable aléatoire asymptotique est chi-carré, centrale si l'hypothèse nulle sous test est vraie, et non centrale sinon. Le réel paramètre de non centralité est une fonction de la mise en marche d'un DGP considéré comme un modèle des différentes possibilités qui existent dans le voisinage de l'hypothèse nulle. Comme les mathématiques impliquées ne sont pas simples, nous n'avons pas discuté des détails expliquant comment ce paramètre non central peut être dérivé, mais l'intuition est essentiellement la même que pour le cas non des modèles de régression non linéaires discutés dans la Section 12.4.

Les propriétés asymptotiques des tests classiques sous des DGP autres que ceux qui satisfassent l'hypothèse nulle sont étudiées dans un article bien connu de Gallant et Holly (1980) aussi bien que dans l'article d'étude de Engle (1984). dans ces articles, seule la mise en marche des DGP qui satisfont l'hypothèse alternative ont été pris en compte. L'article de Gallant et Holly a provoqué une quantité énormes de recherches complémentaires.



Une référence de la littérature dans laquelle cette recherche est rapportée est un article écrit par Burguete, Gallant, et Souza (1982), dans lequel est entrepris un projet ambitieux d'unification d'une large variété de méthodes asymptotiques. Ici, pour la première fois, les mises en marche des DGP qui ont été considérées, bien que dans le voisinage de l'hypothèse nulle, n'ont satisfait ni l'hypothèse nulle, ni l'hypothèse alternative. Plus tard, Newey (1985a) et Tauchen (1985) ont continué le principe de cette approche et ont été amenés à proposer de nouveaux tests et encore plus de procédures de test (voir le Chapitre 16). Notre propre article (Davidson et MacKinnon, 1987) a poursuivi l'étude des locaux généraux DGP et fut parmi les premiers à essayer de construire la théorie du test d'hypothèse dans une charpente géométrique d'une telle manière que les "alentours" de l'hypothèse nulle pourraient être formellement définis et visualisés mentalement. L'approche géométrique avait gagné la faveur des économètres et, plus particulièrement, des statisticiens pour un certain temps avant ceci et avait conduit aux synthèses trouvées dans Amari (1985) et Barndorff-Nielsen, Cox, et Reid (1986); consulter l'article d'étude écrit par Kass (1989). Nous devrions avertir les lecteurs, cependant, que les dernières quelques références citées utilisent des mathématiques qui sont loin d'être élémentaires.

Dans certaines manières, l'approche la plus satisfaisante intuitivement pour le test est fournie par le concept des régressions artificielles. Ce concept, que nous utilisons depuis le Chapitre 6 et que nous développerons plus loin dans la suite de cet ouvrage, fournit, comme des lecteurs l'ont peut-être déjà pressenti, beaucoup d'intuition apportée par une plus haute puissance et analyses sophistiquées mathématiquement. Cela fournit également des moyens simples pour calculer des statistiques de test en pratique.

## TERMES ET CONCEPTS

tests $C(\alpha)$	distribution d'une chi-deux centrale
statistiques classiques de test	régression du produit-extérieur-au-
contradiction parmi les critères de	gradient (OPG)
test	reparamétrisation d'un modèle
forme score efficace de la statistique	paramétrisé
LM	estimations contraintes
produit interne (pour un espace	modèle contraint
Euclidien)	taille $t$ de l'échantillon
invariance (à la reparamétrisation)	vecteur score (vecteur gradient)
terme d'ordre induit (d'un	restrictions lisses
développement asymptotique)	estimations non contraintes
reparamétrisation linéaire	modèles non contraints
alternatives localement équivalente	somme vectorielle
modèles localement équivalents	statistiques tout-comme-Wald
(modèles qui se touchent)	

# Chapitre 14

## Transformation de la Variable Dépendante

### 14.1 INTRODUCTION

Quand nous avons introduit le concept d'une fonction de régression dans le Chapitre 2, nous l'avons défini comme la fonction qui détermine la moyenne d'une variable dépendante  $y_t$  conditionnelle à un ensemble d'information  $\Omega_t$ . Avec cette définition, nous pouvons toujours écrire

$$y_t = x_t(\beta) + u_t \quad (14.01)$$

et affirmer que  $u_t$  a une moyenne nulle conditionnelle à  $\Omega_t$ , à condition que  $x_t(\beta)$  ait été correctement spécifiée. Cependant, quelle que soit la façon correcte dont  $x_t(\beta)$  a été spécifiée, nous ne pouvons pas affirmer que  $u_t$  soit i.i.d. ou possède d'autres propriétés souhaitables. En particulier, il n'y existe aucune raison pour que  $u_t$  soit normalement distribué, homoscédastique, ou même symétrique. Cependant nous avons besoin que  $u_t$  soit homoscédastique pour que les estimations NLS  $\hat{\beta}$  soient efficaces et pour que les inférences basées sur l'estimateur habituel des moindres carrés de la matrice de covariance soient valides.<sup>1</sup> Nous avons aussi besoin que  $u_t$  soit symétrique (et normalement distribué de préférence ou proche de l'être) pour que les résultats asymptotiques fournissent une bonne indication sur les propriétés des estimateurs en échantillon fini. De plus, si nous désirons prédire  $y_t$  conditionnelle à  $\Omega_t$  et construire un quelconque intervalle de prévision, nous devons connaître (ou du moins être capable d'estimer) la distribution de  $u_t$ .

Si nous pouvons trouver la moyenne de  $y_t$  conditionnelle à  $\Omega_t$ , alors nous pouvons probablement tout aussi bien trouver la moyenne conditionnelle de n'importe quelle fonction monotone lisse de  $y_t$ , disons  $\tau(y_t)$ . Par exemple,  $\tau(y_t)$  pourrait être  $\log y_t$ ,  $y_t^{1/2}$ , ou  $y_t^2$ . Si nous écrivons

$$\tau(y_t) = E(\tau(y_t) \mid \Omega_t) + v_t \quad (14.02)$$

<sup>1</sup> Comme nous l'avons vu dans la Section 11.6 et comme nous en discuterons prochainement dans le Chapitre 16, il est possible de réaliser des inférences asymptotiquement valides même en présence d'une forme inconnue d'hétéroscédasticité. Cependant les inférences en échantillon fini seront presque toujours plus précises si les aléas sont homoscédastiques au départ.

pour un quelconque  $\tau(\cdot)$  non linéaire, alors l'aléa  $v_t$  ne peut pas être normalement et indépendamment distribué, ou n.i.d., si  $u_t$  est n.i.d. dans (14.01). Inversement, si  $v_t$  est n.i.d. dans (14.02),  $u_t$  ne peut pas être n.i.d. dans (14.01).

Considérons maintenant un exemple concret, et très réaliste. Supposons que nous estimions le modèle (14.01) quand le DGP pour  $y_t$  est réellement

$$\log y_t = \log(m_t) + v_t, \quad (14.03)$$

où  $m_t$  est dans l'ensemble d'information  $\Omega_t$ , et l'aléa  $v_t$  est  $\text{NID}(0, \sigma^2)$ . Il s'ensuit que

$$y_t = \exp(\log(m_t) + v_t) = m_t \exp(v_t) \cong m_t(1 + v_t) = m_t + m_t v_t,$$

où l'approximation  $\exp(v_t) \cong 1 + v_t$  qui est utilisée ici sera satisfaisante lorsque  $\sigma$  est petit. Si  $m_t = x_t(\beta_0)$  pour un quelconque  $\beta_0$ , la régression non linéaire (14.01) est au moins approximativement valide pour la moyenne conditionnelle de  $y_t$ , bien que ceci ne soit pas nécessairement le cas si la transformation dans (14.03) n'était pas logarithmique. Mais les aléas  $u_t$  qui adhèrent à  $m_t$  ne peuvent pas être n.i.d. En effet, ils seront hétéroscédastiques, avec une variance proportionnelle au carré de  $x_t(\beta_0)$ . Ils seront aussi quelque peu asymétriques à droite, particulièrement si  $\sigma$  n'est pas très petit, parce que le fait que, pour  $a > 0$ ,  $e^a - 1 > |e^{-a} - 1|$ . Ceci implique que toute valeur positive de  $v_t$  se transforme en un  $u_t$  dont la valeur absolue est plus grande que celle apportée par  $-v_t$ . Comme  $v$  est symétrique,  $u$  doit alors être asymétrique à droite.

Cet exemple démontre que, même quand la variable dépendante avait été réellement générée par un DGP à erreurs n.i.d., l'utilisation de la mauvaise transformation de la variable dépendante comme régressande fournira en général une régression à aléas ni homoscedastiques ni symétriques. Ainsi, quand nous rencontrerons l'hétéroscédasticité et l'asymétrie dans les résidus d'une régression, une façon possible de les éliminer consistera à estimer un modèle de régression différent dans lequel la variable dépendante a été soumise à une **transformation non linéaire**. Il s'agit en fait d'une approche déjà grandement utilisée en économétrie et en statistique, et dont nous discutons plus en détail dans ce chapitre. Cependant, nous devrions insister dès à présent que dans n'importe quel cas donné il peut ne pas exister de transformation de la variable dépendante qui fournisse des résidus symétriques et homoscedastiques. Il est également possible qu'une certaine forme des moindres carrés pondérés fonctionnera mieux qu'un modèle qui comporte une transformation de la variable dépendante. Ainsi les techniques qui seront discutées dans ce chapitre ne seront pas utiles pour chaque cas.

Il existe de nombreuses manières où les transformations de la variable dépendante peuvent être employées dans un modèle de régression. Désignons  $\tau(x, \lambda)$  une transformation non linéaire de  $x$  avec comme paramètre scalaire

$\lambda$  qui peut ou pas avoir été estimé. La transformation la plus communément répandue est la **transformée de Box-Cox**, qui a été proposée par Box et Cox (1964) dans un très célèbre article; elle sera discutée dans la prochaine section. Une classe de modèle qui utilise une telle transformation est celle suggérée à l'origine par Box et Cox:

$$\tau(y_t, \lambda) = x_t(\boldsymbol{\beta}) + u_t, \quad (14.04)$$

où la transformée s'applique seulement à la variable dépendante. Cette classe de modèles a été très courante en statistique mais beaucoup moins en économétrie. Une seconde classe de modèles est

$$\tau(y_t, \lambda) = \tau(x_t(\boldsymbol{\beta}), \lambda) + u_t, \quad (14.05)$$

dans laquelle la transformation  $\tau(x, \lambda)$  est appliquée à la fois à la variable et à la fonction de régression. Les modèles de ce type avaient été préconisés par Carroll et Ruppert (1984, 1988), qui les ont appelés modèles de “transformation des deux cotés”. Ces modèles ont aussi été très largement utilisés en statistique et dans une moindre mesure en économétrie; un exemple précoce est Leech (1975).

Une troisième classe de modèles est

$$\tau(y_t, \lambda) = \sum_{i=1}^k \beta_i \tau(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad (14.06)$$

où  $X_{ti}$  et  $Z_{tj}$  désignent à la fois les observations sur les variables indépendantes, la distinction étant que les  $X_{ti}$  sont soumis à la transformée et que les  $Z_{tj}$  ne le sont pas. Il s'agit de l'approche qui a été généralement prise en économétrie, avec la transformation  $\tau(x, \lambda)$  étant immanquablement la transformée de Box-Cox.<sup>2</sup> La classe de modèles (14.06) est plus générale que (14.04), au moins si  $x_t(\boldsymbol{\beta})$  dans ce modèle est restreint à être linéaire, et d'une certaine manière elle est aussi plus générale que (14.05). Elle peut aussi être généralisée par la suite en permettant à la valeur de  $\lambda$  utilisée pour transformer  $y_t$  d'être différent de la valeur (ou des valeurs) utilisée pour transformer les  $X_{ti}$  (consulter la Section 14.7).

Notons que les modèles (14.04) et (14.05) sont principalement concernés par l'obtention de résidus homoscédastiques et symétriques, alors que la forme fonctionnelle de la fonction de régression est considérée comme une donnée. En revanche, dans le modèle (14.06), la forme fonctionnelle dépend explicitement de  $\lambda$ . Peut-être qu'en conséquence de ceci, la majeure partie de la littérature des débuts de l'économétrie s'est principalement intéressée

<sup>2</sup> Les études qui utilisent ou discutent cette approche incluent Zarembka(1968, 1974), White (1972), Heckman et Polachek (1974), Savin et White (1978), et Spitzer (1976, 1978, 1982a, 1982b, 1984).

à déterminer la forme fonctionnelle de la fonction de régression, en restant très peu concernée par les propriétés des résidus. Ce manque d'intérêt a été mal placé, parce que la caractéristique clé de tout modèle comprenant une transformation de la variable dépendante est que la transformation affecte directement les propriétés des résidus.

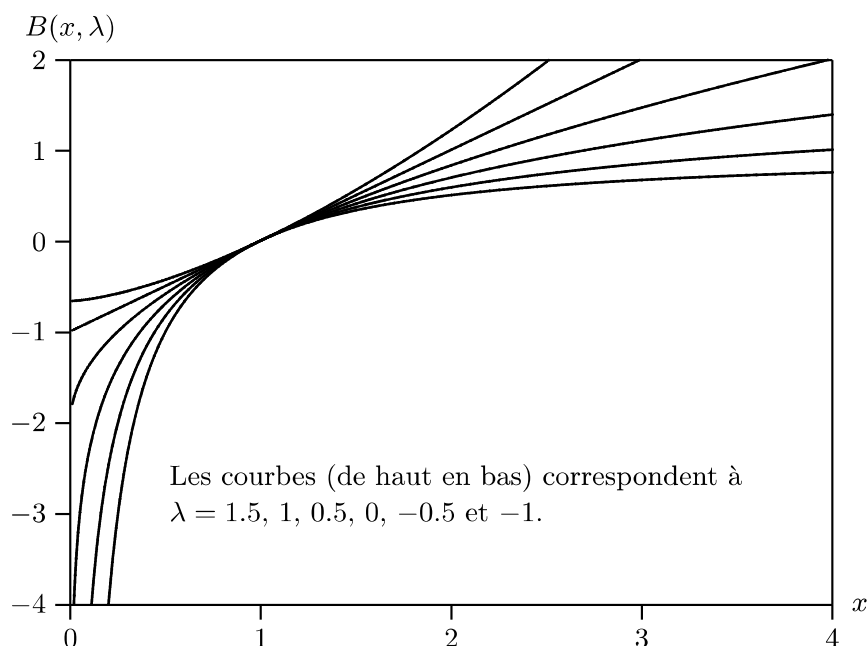
Les modèles (14.04), (14.05), et (14.06) ne peuvent pas être qualifiés de modèles de régression, parce que la variable dépendante n'est pas simplement égale à la somme d'une fonction de régression et d'un aléa. Bien que ces modèles soient différents, et puissent fournir des résultats très différents en pratique, ils comportent tous un élément en commun, à savoir, que la variable dépendante est soumise à une transformation non linéaire avec le paramètre  $\lambda$ . Si  $\lambda$  était connu, ces modèles pourraient tous être estimés par moindres carrés non linéaires et testés en utilisant la régression de Gauss-Newton. Mais tant que  $\lambda$  est inconnu et qu'il doit être estimé, la méthode NLS est clairement inappropriée. Dans la plupart des cas, un algorithme par moindres carrés choisirait simplement  $\lambda$  de façon à rendre  $\tau(y_t, \lambda)$  aussi petit que possible afin de rendre la somme des résidus au carré aussi petite que possible. Ainsi, cela fournirait inévitablement des résultats insensés, comme nous en avons discuté dans le Chapitre 8 en connexion avec le modèle (8.01).

Dans la prochaine section, nous discutons de la transformée de Box-Cox et de l'estimation des modèles de régression où la variable dépendante a été soumise à cette transformation. L'estimation par maximum de vraisemblance se trouve être très facile parce que la fonction de logvraisemblance incorpore un terme Jacobien qui empêche  $\lambda$  de devenir trop petit. Dans la Section 14.3, nous réalisons une légère digression pour discuter de certaines des propriétés utiles des termes Jacobiens dans l'estimation ML. Dans la section 14.4, nous discutons alors d'une nouvelle classe de régressions artificielles appelée **régressions artificielles à longueur double** et, dans la Section 14.5, nous montrons comment celles-ci peuvent être utilisées pour l'estimation et le test des modèles comprenant la transformée de Box-Cox. Dans la Section 14.6, nous discutons de la façon dont il est possible de tester la spécification de linéarité ou de loglinéarité d'un modèle contre une alternative Box-Cox ou autre. Finalement, dans la Section 14.7, nous traitons brièvement de certains modèles qui comprennent des généralisations ou des alternatives de la transformée de Box-Cox.

## 14.2 LA TRANSFORMÉE DE BOX-COX

La transformée de Box-Cox est la transformation non linéaire de loin la plus rencontrée en statistique et en économétrie. Elle est définie comme

$$B(x, \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{quand } \lambda \neq 0; \\ \log(x) & \text{quand } \lambda = 0, \end{cases}$$



**Figure 14.1** Transformées de Box-Cox pour des valeurs diverses de  $\lambda$

où l'argument  $x$  doit être positif. D'après la règle de l'Hôpital,  $\log x$  est la limite de  $(x^\lambda - 1)/\lambda$  quand  $\lambda \rightarrow 0$ . La Figure 14.1 montre la transformée de Box-Cox pour différentes valeurs de  $\lambda$ . En pratique,  $\lambda$  s'étend généralement d'une valeur inférieure à 0 à une valeur supérieure à 1. Il peut être montré que  $B(x, \lambda') \geq B(x, \lambda'')$  pour  $\lambda' \geq \lambda''$ , et cette inégalité est évidente sur la figure. Ainsi la valeur de courbure de la transformée de Box-Cox augmente quand  $\lambda$  s'éloigne de 1 dans l'une ou l'autre direction.

Il existe trois variétés de modèle de Box-Cox. Nous nous référerons à (14.04) et (14.05) avec  $\tau(\cdot)$  donné par la transformée de Box-Cox, du **modèle de Box-Cox simple** et du **modèle de Box-Cox transformé des deux côtés** respectivement. Nous nous référerons à (14.06) du **modèle de Box-Cox conventionnel**, parce qu'il s'agit du plus communément utilisé en économétrie.

Une des raisons de la popularité de la transformée de Box-Cox est qu'elle incorpore à la fois la possibilité d'aucune transformation (quand  $\lambda = 1$ ) et la possibilité d'une transformation logarithmique (quand  $\lambda = 0$ ). Sous réserve que les régresseurs incluent un terme constant, soumettre la variable dépendante à la transformée de Box-Cox  $\lambda = 1$  est équivalent à n'effectuer aucune transformation. Soumettre la variable dépendante à la transformée de Box-Cox avec  $\lambda = 0$  est équivalent à utiliser  $\log y_t$  comme régressande. Comme ces deux transformations sont deux cas spécifiques très plausibles, il est très séduisant d'utiliser une transformation qui tienne compte des deux à la fois. Même quand le modèle conventionnel de Cox-Box n'est pas considéré comme véritablement plausible, ce dernier fournit une alternative commode

à partir de laquelle il est possible de tester la spécification des modèles de régression linéaire et non linéaire; consulter la Section 14.6.

Cependant, la transformée de Box-Cox n'est pas sans sérieux inconvénients. Considérons le modèle de Box-Cox simple

$$B(y_t, \lambda) = x_t(\beta) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \quad (14.07)$$

Pour la plupart des valeurs de  $\lambda$  (mais pas pour  $\lambda = 0$  ou  $\lambda = 1$ ) la valeur de  $B(y_t, \lambda)$  est bornée soit inférieurement soit supérieurement; de manière spécifique, quand  $\lambda > 0$ ,  $B(y_t, \lambda)$  ne peut pas être inférieur à  $-1/\lambda$  et, quand  $\lambda < 0$ ,  $B(y_t, \lambda)$  ne peut pas être supérieur à  $-1/\lambda$ . Cependant, si  $u_t$  est normalement distribué, le membre de droite de (14.07) n'est pas borné et pourrait, du moins en principe, prendre des valeurs positives ou négatives arbitrairement grandes. Ainsi, à strictement parler, (14.07) est logiquement impossible en tant que modèle pour  $y_t$ . Ceci reste vrai si nous remplaçons  $x_t(\beta)$  par une fonction de régression qui dépend de  $\lambda$ .

Une manière de traiter ce problème est de supposer que les données sur  $y_t$  sont observées seulement quand les bornes ne sont pas enfreintes, comme dans Poirier (1978) et Poirier et Ruud (1979). Ceci nous conduit à discuter des fonctions de logvraisemblance similaires à celles discutées dans la Section 15.6.<sup>3</sup> Cependant, rien ne justifie le fait que les données doivent toujours être générées de cette manière, et de plus, à la fois l'estimation et le test deviennent très compliqués quand on prend en compte cette sorte de troncature d'échantillon. Une seconde manière de traiter de ce problème consiste simplement à l'ignorer. Cette application du bien célèbre "algorithme autruche" prend tout son sens si  $\lambda$  est non négatif (ou du moins pas inférieur à zéro) et  $y_t$  est positive est relativement grande par rapport à  $\sigma$  pour toutes les observations dans l'ensemble d'information. Quand ces deux conditions sont satisfaites, nous pouvons être sûrs que  $u_t$  sera plus petit relativement à  $B(y_t, \lambda)$  et  $x_t(\beta)$ ; par conséquent, la probabilité que le membre de droite de (14.07) n'enfreigne la borne du membre de gauche sera très petite.

Nous adopterons cette seconde approche, parce que les modèles de Box-Cox comportant des valeurs négatives de  $\lambda$  ne sont pas très intéressants, et que, dans de nombreux cas pratiques, la moyenne conditionnelle de  $y_t$  est toujours relativement grande par rapport à n'importe quelle variation autour de la moyenne conditionnelle. Dans de tels cas, il semble assez raisonnable d'utiliser des modèles dans lesquels la variable dépendante est soumise à la transformée de Box-Cox. Cependant, dans d'autres cas, il peut ne pas être correct d'utiliser un modèle de Box-Cox; consulter la Section 14.7.

Considérons maintenant la façon d'obtenir des estimations convergentes de  $\lambda$  et  $\beta$  dans (14.07). Il s'agit du cas le plus simple à discuter, mais tout ce

<sup>3</sup> Une approche différente, mais similaire, avait été proposée par Amemiya et Powell (1981).

que nous dirons s'appliquera également, avec quelques légères et évidentes modifications, aussi bien aux modèles transformés des deux côtés qu'aux modèles de Box-Cox conventionnels, dans lesquels le paramètre de transformation  $\lambda$  apparaît également dans la fonction de régression. Puisque, clairement, les moindres carrés ne serviront pas dans ce cas, il est naturel de se tourner vers le maximum de vraisemblance. Comme nous avons supposé que les  $u_t$  sont normalement et indépendamment distribués, nous pouvons facilement écrire la fonction de logvraisemblance pour ce modèle. Il s'agit de

$$\begin{aligned} \ell(\mathbf{y}, \boldsymbol{\beta}, \lambda, \sigma) = & -\frac{n}{2} \log(2\pi) - n \log \sigma \\ & - \frac{1}{2\sigma^2} \sum_{t=1}^n (B(y_t, \lambda) - x_t(\boldsymbol{\beta}))^2 + (\lambda - 1) \sum_{t=1}^n \log y_t. \end{aligned} \quad (14.08)$$

Ici le dernier terme est la somme sur toutes les observations du logarithme de

$$\frac{\partial B(y_t, \lambda)}{\partial y_t} = \frac{\partial}{\partial y_t} \left( \frac{y_t^\lambda - 1}{\lambda} \right) = y_t^{\lambda-1},$$

qui est le Jacobien de la transformation de  $y_t$  vers  $u_t$ .

Le rôle de ce terme Jacobien est crucial. Afin d'éviter d'avoir plus d'un cas, supposons pour simplifier que tous les  $y_t$  sont plus grands que 1. Puisque

$$\lim_{\lambda \rightarrow -\infty} B(x, \lambda) = 0$$

pour  $x > 1$ , en laissant  $\lambda \rightarrow -\infty$  alors  $B(y_t, \lambda) \rightarrow 0$  pour tout  $t$ . Ainsi, à condition qu'il existe une certaine valeur de  $\boldsymbol{\beta}$  qui fait que la fonction de régression  $x_t(\boldsymbol{\beta})$  égale zéro pour tout  $t$ , la somme des résidus au carré,

$$\sum_{t=1}^n (B(y_t, \lambda) - x_t(\boldsymbol{\beta}))^2,$$

devient arbitrairement petite, si on laisse  $\lambda$  tendre vers moins l'infini. Si nous concentrons (14.08) par rapport à  $\sigma$ , la fonction de logvraisemblance devient

$$\ell^c(\mathbf{y}, \boldsymbol{\beta}, \lambda) = C - \frac{n}{2} \log \left( \sum_{t=1}^n (B(y_t, \lambda) - x_t(\boldsymbol{\beta}))^2 \right) + (\lambda - 1) \sum_{t=1}^n \log y_t, \quad (14.09)$$

où  $C$  est une constante qui ne dépend ni de  $\boldsymbol{\beta}$  ni de  $\lambda$ . Ainsi, nous voyons que lorsque nous maximisons la fonction de logvraisemblance, la valeur de  $\lambda$  affectera deux éléments: un terme somme-des-carrés et un terme Jacobien. Le terme Jacobien empêche l'estimation ML de  $\lambda$  de tendre vers moins l'infini, puisque ce terme tend vers moins l'infini tout comme  $\lambda$ .



La maximisation de (14.09) n'est pas très difficile. La meilleure approche, si le logiciel approprié est commode, consiste à utiliser une procédure convenable pour la maximisation non linéaire; consulter la Section 14.5. Une seconde approche consiste à employer une procédure par balayage dans laquelle on recherche les valeurs de  $\lambda$  et estime  $\beta$  par moindres carrés conditionnellement à  $\lambda$ . Une troisième approche consiste à utiliser une astuce qui permet à (14.09) d'être minimisée en utilisant n'importe quel algorithme de moindres carrés non linéaires. Il existe en fait deux manières de réaliser ceci. La plus simple est de noter que si tous les  $y_t$  sont divisés par leur moyenne géométrique  $\dot{y}$ , le terme Jacobien dans (14.09) est alors identiquement égal à zéro, parce que

$$n \log \dot{y} = \sum_{t=1}^n \log y_t.$$

Ainsi, n'importe quelle régression dont les résidus sont  $B(y_t/\dot{y}, \lambda) - x_t(\beta)$  fournira des estimations valides de  $\beta$  et  $\lambda$ . Par exemple, nous pourrions définir la régressande comme un vecteur de zéros et la fonction de régression comme  $B(y_t/\dot{y}, \lambda) - x_t(\beta)$  et alors utiliser n'importe quel algorithme. Cette approche a été usitée pendant de nombreuses années mais comporte le désavantage qu'il faut modifier l'échelle des  $y_t$ ; comme nous le verrons plus loin, cette procédure n'est pas toujours totalement neutre dans le contexte des modèles de Box-Cox.

Une seconde manière d'utiliser un programme NLS a été proposée par Carroll et Ruppert (1988). Nous pouvons récrire (14.09) comme

$$\ell^c(\mathbf{y}, \beta, \lambda) = C^* - \frac{n}{2} \log \left( \sum_{t=1}^n \left( \frac{B(y_t, \lambda) - x_t(\beta)}{\dot{y}^\lambda} \right)^2 \right),$$

où  $C^*$  ne dépend ni de  $\beta$  ni de  $\lambda$ . Comme cette version de la fonction de logvraisemblance n'a seulement qu'un terme en somme-des-carrés, elle peut être maximisée par la minimisation de la somme des résidus au carré:

$$\sum_{t=1}^n \left( \frac{B(y_t, \lambda) - x_t(\beta)}{\dot{y}^\lambda} \right)^2.$$

Nous pouvons procéder ainsi en utilisant une procédure NLS en définissant la régressande comme un vecteur de zéros et la fonction de régression comme  $(B(y_t, \lambda) - x_t(\beta))/\dot{y}^\lambda$ .

Bien que toutes les techniques précédemment décrites fournissent des estimations ML  $\hat{\lambda}$  et  $\hat{\beta}$ , aucune des méthodes basées sur les moindres carrés ne fournit une estimation valide de la matrice de covariance  $\hat{\lambda}$  et  $\hat{\beta}$ . La raison est, comme nous le verrons dans la Section 14.5, la matrice d'information pour les modèles de Box-Cox n'est pas bloc-diagonale en  $\beta$ ,  $\lambda$ , et  $\sigma$ . Les méthodes de grille de valeurs qui estiment  $\beta$  conditionnellement à  $\lambda$  fournissent des matrices de covariance invalides parce qu'elles ignorent le fait que  $\hat{\lambda}$  est elle-même une

estimation. Les méthodes qui amènent un programme NLS à estimer  $\hat{\lambda}$  et  $\hat{\beta}$  conjointement fournissent également des estimations de matrices de covariance invalides parce qu'elles supposent implicitement que la matrice de covariance est bloc-diagonale entre  $\sigma$  et les autres paramètres, ce qui n'est pas le cas pour les modèles de Box-Cox. Comme il est très tentant d'utiliser les estimations incorrectes de l'écart type affichées par le progiciel des moindres carrés, nous recommandons que les procédures basées sur moindres carrés soient seulement usitées pour estimer les modèles de Box-Cox quand le logiciel le plus approprié est indisponible.

Nous pouvons, naturellement, obtenir une matrice de covariance estimée valide de différentes façons en inversant différentes estimations de la matrice d'information. La régression OPG fournit probablement la manière la plus simple d'obtenir une estimation de matrice de covariance, mais ses propriétés en échantillon fini ne sont pas très bonnes, et des techniques plus spécialisées plus appropriées sont disponibles; consulter Spitzer (1984). Dans la Section 14.4 et 14.5, nous discuterons d'une classe de régressions artificielles qui peut être utilisée pour traiter d'une large classe de modèles et semble très bien fonctionner pour les modèles de Box-Cox. Comme toutes les régressions artificielles, ces régressions à longueur double, tel est leur nom, peuvent être employées pour des estimations, inférences, et tests de spécification.

Nous avons remarqué plus tôt que la modification d'échelle de la variable dépendante peut ne pas être neutre dans un modèle de Box-Cox. Dans un modèle transformé des deux côtés, la renormalisation de la variable dépendante a exactement le même effet que s'il n'y avait eu aucune transformation, parce qu'à la fois la variable dépendante et la fonction de régression sont transformées de la même façon. Ainsi, si  $x_t(\beta)$  est linéaire, tous les coefficients seront simplement multipliés par le facteur utilisé pour renormaliser la variable dépendante. Si  $x_t(\beta)$  est non linéaire, la renormalisation de  $y_t$  peut très bien avoir une influence sur  $\beta$  de façon plus délicate et peut même affecter la façon dont le modèle s'ajuste, mais cela se réalisera seulement si la renormalisation affecte l'ajustement du modèle même si aucune transformation n'est impliquée. Cependant, dans les deux autres types de modèle de Box-Cox, les choses ne sont pas aussi simples.

Il existe un important résultat d'invariance pour les modèles de Box-Cox conventionnels et simples. Ce résultat est que, sous certaines conditions, l'estimation de  $\lambda$  est invariante par rapport à l'échelle de la variable dépendante. Supposons que nous multiplions  $y_t$  par une constante  $\alpha$  de sorte que la variable dépendante devienne  $\alpha y_t$ . La transformée de Box-Cox de  $\alpha y_t$  est

$$B(\alpha y_t, \lambda) = \alpha^\lambda B(y_t, \lambda) + B(\alpha, \lambda).$$

Ici le second terme est juste une constante. Pourvu qu'il y ait un terme constant (ou l'équivalent) dans la fonction de régression, l'estimation de la constante s'ajustera toujours automatiquement pour s'y accommoder. Si la fonction de régression est linéaire, toutes les estimations paramétriques sauf

la constante seront simplement multipliées par  $\alpha^\lambda$ , tout comme les résidus et  $\hat{\sigma}$ . Pour le modèle de Box-Cox, la renormalisation est plus compliquée, mais l'effet net est que les résidus sont encore multipliés par  $\alpha^\lambda$ . Ceci est également vrai pour certaines autres fonctions de régression  $x_t(\beta)$ , mais pas pour toutes. Pourvu que la renormalisation  $y_t$  soit équivalente à celle des résidus de cette manière, le terme somme-des-carrés dans (14.08), évalué pour un  $\lambda$  fixé arbitrairement au  $\hat{\beta}$  qui minimise la somme des résidus au carré et le correspondant  $\hat{\sigma}^2$ , est invariant sous la renormalisation. Le second terme de (14.08),  $-n \log \sigma$ , devient  $-n \log \sigma - n\lambda \log \alpha$ . Le dernier terme, le Jacobien, devient

$$(\lambda - 1) \sum_{t=1}^n \log y_t + n(\lambda - 1) \log \alpha.$$

Ainsi l'opération entière additionne  $-n \log \alpha$ , une quantité indépendante de tous les autres paramètres, à la fonction de logvraisemblance concentrée par rapport à  $\beta$  et  $\sigma^2$ . Par conséquent il est clair que, pourvu que la normalisation de  $y_t$  est équivalente à celle des résidus, l'estimation ML  $\hat{\lambda}$  ne changera pas quand nous renormalisons  $y_t$ . Ce résultat a été, pour l'essentiel, démontré à l'origine par Schlesselman (1971).

Même quand  $\hat{\lambda}$  est invariant à l'échelle, les autres paramètres ne le seront généralement pas. Dans le modèle de Box-Cox conventionnel, les effets de la renormalisation de  $y_t$  dépendent de la valeur de  $\lambda$ . Quand  $\lambda = 1$ , de telle sorte qu'il s'agisse réellement d'un modèle de régression linéaire, la multiplication de  $y_t$  par  $\alpha$  change simplement tous les coefficients estimés par un facteur de  $\alpha$  et n'a aucun effet sur les  $t$  de Student. Quand  $\lambda = 0$ , de telle sorte qu'il s'agisse réellement d'un modèle de régression loglinéaire, la multiplication de  $y_t$  par  $\alpha$  signifie l'addition d'un  $\log \alpha$  constant à la régressande, qui affecte le terme constant mais aucun des autres coefficients. Mais à l'exception de ces deux cas, tous les autres coefficients changeront généralement quand la variable dépendante est renormalisée. De plus, en raison du manque d'invariance des tests de Wald aux reparamétrisations non linéaires, tous les  $t$  de Student sur les  $\beta_i$  changeront de la sorte; consulter Spitzer (1984). En fait, il est très possible qu'un Student soit hautement significatif pour une normalisation de  $y_t$  et complètement non significatif pour une autre. Ceci implique naturellement que, quelle que soit la normalisation de  $y_t$ , nous ne devons pas faire confiance aux Student (ou à n'importe quelle sorte test de Wald) dans le contexte des modèles de Box-Cox.

### 14.3 LE RÔLE DES TERMES JACOBIENS DANS L'ESTIMATION ML

Des termes Jacobiens sont apparus dans les fonctions de logvraisemblance dans une variété de contextes dans les Chapitres 8, 9, et 10. Nous avons vu qu'à chaque fois que la variable dépendante est soumise à une transformation non linéaire, la fonction de logvraisemblance contient nécessairement au moins

un terme Jacobien. Dans cette section, nous étudions plus en détails le rôle joué par les termes Jacobiens dans l'estimation ML. Nous continuerons notre discussion des modèles de Box-Cox dans les sections suivantes.

Rappelons que si la densité de probabilité d'une variable aléatoire  $x_1$  est  $f_1(x_1)$  et qu'une autre variable aléatoire  $x_2$  lui y est reliée par  $x_1 = \tau(x_2)$ , où la fonction  $\tau(\cdot)$  est continuellement différentiable et monotone, alors la densité de  $x_2$  est donnée par

$$f_2(x_2) = f_1(\tau(x_2)) \left| \frac{\partial \tau(x_2)}{\partial x_2} \right|. \quad (14.10)$$

Le second facteur ici est la valeur absolue du Jacobien de la transformation, et est alors souvent désigné sous le nom de **facteur Jacobien**. Dans le cas multivarié, où  $\mathbf{x}_1$  et  $\mathbf{x}_2$  sont des vecteurs de dimension  $m$  et  $\mathbf{x}_1 = \boldsymbol{\tau}(\mathbf{x}_2)$ , l'analogue de (14.10) est

$$f_2(\mathbf{x}_2) = f_1(\boldsymbol{\tau}(\mathbf{x}_2)) |\det \mathbf{J}(\mathbf{x}_2)|,$$

où  $|\det \mathbf{J}(\mathbf{x}_2)|$  est la valeur absolue du déterminant de la matrice Jacobienne  $\mathbf{J}(\mathbf{x}_2)$  avec comme élément type

$$J_{ij}(\mathbf{x}_2) \equiv \frac{\partial \tau_i(\mathbf{x}_2)}{\partial x_{2j}}.$$

Ces résultats sont discutés dans l'Annexe B.

Des facteurs Jacobiens dans les fonctions de densité induisent des termes Jacobiens dans les fonctions de logvraisemblance. Ceux-ci peuvent se produire à chaque fois que la transformation de la variable(s) dépendante(s) observée(s) vers les aléas comporte une matrice Jacobienne qui n'est pas la matrice identité. Si les aléas sous-jacents sont supposés être normalement distribués, la présence de ces termes Jacobiens est souvent la seule chose qui fait que la fonction de logvraisemblance soit autre chose qu'une banale transformation de la somme des résidus au carré.

Cependant, il existe des circonstances dans lesquelles la fonction de logvraisemblance ne contient aucun terme Jacobien, même si la matrice Jacobienne n'est pas une matrice identité. Nous avons rencontré une classe de modèles pour lesquels ceci est le cas dans le Chapitre 10. Si nous oublions la première observation, la matrice Jacobienne pour un modèle de régression à erreurs AR(1) est triangulaire inférieure, avec des éléments diagonaux égaux à 1. Puisque le déterminant d'une matrice triangulaire est le produit des éléments de la diagonale, le facteur Jacobien pour de tels modèles est simplement l'unité, et le terme Jacobien est par conséquent zéro.

Dans cette section, naturellement, nous traitons de nombreux autres cas dans lesquels les termes Jacobiens apparaissent dans des fonctions de logvraisemblance. Leur apparition comporte plusieurs conséquences. Tout

d'abord, elle signifie que les moindres carrés non linéaires et la régression de Gauss-Newton ne sont pas applicables à de tels modèles. Des astuces telles que celle que nous avons utilisée dans la section précédente peuvent permettre aux NLS d'être utilisées pour l'estimation, mais elles ne permettront pas à l'inférence d'être basée comme d'habitude sur les estimations NLS. La régression OPG sera applicable, ainsi que des régressions artificielles plus spécialisées telles que la régression à longueur double qui sera introduite dans la prochaine section.

Ensuite, la présence des termes Jacobiens assure que nous ne pouvons jamais obtenir des estimations en des points dans l'espace paramétrique où le Jacobien de la transformation provenant de la variable(s) dépendante(s) vers les aléas sous-jacents est singulier. En de tels points, il ne sera pas du tout possible de réaliser cette transformation. Quand le vecteur de paramètres se rapproche d'un tel point, le déterminant de la matrice Jacobienne tend vers zéro, et le logarithme de ce déterminant tend par conséquent vers moins l'infini. Nous avons vu un exemple de ce phénomène dans la Section 10.6, où la fonction de logvraisemblance pour un modèle à erreurs AR(1) tendait vers moins l'infini quand  $|\rho| \rightarrow 1$ . La transformation pour la première observation,

$$(1 - \rho^2)^{1/2}(y_1 - x_1(\beta)) = \varepsilon_1,$$

ne peut pas être effectuée lorsque  $|\rho| = 1$ , et la fonction de logvraisemblance reflète ce fait en prenant la valeur de moins l'infini.

Cette propriété des fonctions de logvraisemblance est une des plus désirables, parce qu'elle nous empêche d'obtenir des estimations insensées. Cependant, elle implique que les fonctions de logvraisemblance pour de tels modèles doivent avoir des maxima multiples. Par exemple, dans le cas le plus simple dans lequel la singularité divise l'espace paramétrique en deux régions, il doit y avoir au moins un maximum dans chacune de ces régions. Ainsi, si nous commençons l'erreur de débiter l'algorithme de maximisation dans la mauvaise région, l'algorithme peut bien manquer de croiser la singularité et ainsi nous trouverons un maximum local qui n'est pas le maximum global; voir MacKinnon (1979). Nous rencontrerons des exemples complémentaires de singularités dans les fonctions de logvraisemblance dans le Chapitre 18 quand nous discuterons de l'utilisation du maximum de vraisemblance pour l'estimation des modèles à équations simultanées.

La troisième conséquence majeure de la présence de termes Jacobiens dans les fonctions de logvraisemblance, et une des plus intéressantes pour nous dans ce chapitre, est que l'estimation par maximum de vraisemblance, à la différence des moindres carrés, n'est pas gênée par les transformations de la variable dépendante, parce que, comme nous l'avons vu dans la dernière section, la présence d'une transformation occasionne la présence d'un terme Jacobien dans la fonction de logvraisemblance. Un problème courant dans les travaux économétriques appliqués consiste à déterminer la transformation la

plus appropriée de la variable dépendante. Par exemple, la théorie économique pourrait admettre les trois spécifications suivantes:

$$H_1: y_t = \alpha_1 + \beta_1 x_t + u_t, \quad (14.11)$$

$$H_2: \log y_t = \alpha_2 + \beta_2 \log x_t + u_t, \text{ et} \quad (14.12)$$

$$H_3: \frac{y_t}{z_t} = \alpha_3 \frac{1}{z_t} + \beta_3 \frac{x_t}{z_t} + u_t, \quad (14.13)$$

où  $z_t$  et  $x_t$  sont des observations sur des variables exogènes ou prédéterminées. Ici, les fonctions de régression sont délibérément très simples, parce que la manière dont elles sont spécifiées est hors de propos du principal argument.

Il n'est clairement pas approprié de comparer les sommes des résidus au carré ou les  $R^2$  issus de (14.11), (14.12), et (14.13). Néanmoins, si nous voulons supposer la normalité, il est très facile de comparer les valeurs des fonctions de logvraisemblance provenant des trois modèles en compétition. Ces fonctions de logvraisemblance, concentrées par rapport au paramètre de variance, sont, respectivement,

$$\ell_1^c(\mathbf{y}, \beta_1) = C - \frac{n}{2} \log \left( \sum_{t=1}^n (y_t - \alpha_1 - \beta_1 x_t)^2 \right), \quad (14.14)$$

$$\ell_2^c(\mathbf{y}, \beta_2) = C - \frac{n}{2} \log \left( \sum_{t=1}^n (\log y_t - \alpha_2 - \beta_2 \log x_t)^2 \right) - \sum_{t=1}^n \log y_t, \quad (14.15)$$

et

$$\ell_3^c(\mathbf{y}, \beta_3) = C - \frac{n}{2} \log \left( \sum_{t=1}^n \left( \frac{y_t}{z_t} - \alpha_3 \frac{1}{z_t} - \beta_3 \frac{x_t}{z_t} \right)^2 \right) - \sum_{t=1}^n \log z_t, \quad (14.16)$$

où la constante  $C$  est la même pour les trois spécifications.

Ce qui rend possible la comparaison de ces trois fonctions de logvraisemblance est la présence des termes Jacobiens dans (14.15) et (14.16). Ils surviennent parce que

$$\frac{\partial \log y_t}{\partial y_t} = \frac{1}{y_t} \quad \text{et} \quad \frac{\partial (y_t/z_t)}{\partial y_t} = \frac{1}{z_t}.$$

Ainsi, si nous souhaitons décider lequel de (14.11), (14.12), et (14.13) s'ajuste le mieux, nous avons simplement à estimer chacun d'entre eux par NLS (ou peut-être par OLS), à retrouver les valeurs des fonctions de logvraisemblance données par le progiciel de régression, à soustraire  $\sum \log y_t$  dans le cas de (14.12) et  $\sum \log z_t$  dans le cas de (14.13), et à comparer les valeurs qui résultent de  $\ell_1$ ,  $\ell_2$ , et  $\ell_3$ . Notons que, pour la plupart des progiciels de régression, les valeurs de  $\ell$  pour (14.12) et (14.13) seront incorrectes quand  $y_t$

(plutôt que  $\log y_t$  ou  $y_t/z_t$ ) est vraiment la variable dépendante. Comme le progiciel ne sait pas que la régressande a été soumise à une transformation, les valeurs qu'il donne omettront les termes Jacobiens dans (14.15) et (14.16).

Cette sorte de procédure peut en fait être réalisée pour tester, et peut-être rejeter, un ou plusieurs des modèles en compétition. Il est possible de voir que chaque paire de  $H_1$ ,  $H_2$ , et  $H_3$  peut être imbriquée dans un modèle plus général comprenant un paramètre supplémentaire. Par exemple, le modèle

$$\frac{y_t}{z_t^\phi} = \alpha \frac{1}{z_t^\phi} + \beta \frac{x_t}{z_t^\phi} + u_t$$

se réduit à  $H_1$  quand  $\phi = 0$  et à  $H_3$  quand  $\phi = 1$ . De façon similaire, le modèle de Box-Cox

$$B(y_t, \lambda) = \alpha + \beta B(x_t, \lambda) + u_t \quad (14.17)$$

se réduit à  $H_1$  quand  $\lambda = 1$  et à  $H_2$  quand  $\lambda = 0$ . Supposons que nous estimions  $H_1$  et  $H_2$ , et que les valeurs de  $\ell_1$  et  $\ell_2$  soient  $-523.4$  et  $-520.7$ , respectivement. Puisque nous connaissons que le modèle emboîtant (14.17) doit s'ajuster au moins aussi bien que celui de  $H_1$  et  $H_2$  qui s'ajuste le mieux, le maximum non contraint de la fonction de logvraisemblance doit être supérieur ou égal à  $-520.7$ . Ainsi une statistique de test LR de  $H_1$  contre le modèle emboîtant doit être supérieure à

$$2(-520.7 - (-523.4)) = 2(523.4 - 520.7) = 5.4.$$

Puisque 5.4 excède la valeur critique à 5% pour un test à un degré de liberté, nous pouvons conclure que le modèle linéaire  $H_1$  sera rejeté à un niveau inférieur à 5% s'il est testé contre le modèle emboîtant, même si nous n'avons pas estimé le dernier ou calculé une statistique de test formelle.

Cet exemple illustre une caractéristique des tests LR qui peut être très commode, à savoir, que nous pouvons parfois mettre une borne inférieure à la statistique de test LR sans réellement estimer le modèle non contraint. Cette caractéristique a été notée par Sargan (1964) dans le contexte du choix entre modèles linéaire et non linéaire; elle est très largement utilisée dans les travaux appliqués, et elle a récemment été proposée comme une base pour la sélection de modèles par Pollak et Wales (1991). La procédure fonctionne seulement dans une direction, naturellement. Ainsi, le fait qu'une bonne performance de  $H_2$  nous permette de rejeter  $H_1$  dans cet exemple ne nous dit rien concernant  $H_2$ .  $H_2$  pourrait très bien être rejetée également si nous l'avions en fait testée contre le modèle emboîtant (consulter la Section 14.6).

## 14.4 RÉGRESSIONS ARTIFICIELLES À LONGUEUR DOUBLE

Pour tous les modèles discutés dans les Sections 14.1 et 14.2, la fonction de logvraisemblance est égale à une somme de contributions pour chacune des  $n$  observations; (14.08) fournit un exemple. Ainsi, la régression OPG pourrait clairement être utilisée pour l'estimation et le test de ces modèles. Cependant, étant donné la performance généralement pauvre en échantillon fini des quantités calculées au moyen de la régression OPG, nous ne préférons pas y baser les inférences. Heureusement, une autre régression artificielle est disponible. Appelée la **régression artificielle à longueur double**, ou **DLR**, elle aussi peut être utilisée avec ces modèles et elle fonctionne vraiment beaucoup mieux que la régression OPG en échantillons finis. Dans cette section, nous fournirons une brève introduction à la DLR. Dans la prochaine section, nous montrons comment elle peut être utilisée dans l'estimation et le test des modèles de Box-Cox. Les principales références à ce sujet sont Davidson et MacKinnon (1984a, 1988). Davidson et MacKinnon (1983a, 1985c), Bera et McKenzie (1986), Godfrey, McAleer, et McKenzie (1988), et MacKinnon et Magee (1990) fournissent des évidences Monte Carlo qui suggèrent que les tests basés sur la DLR ont des performances bien meilleures qu'en ont ceux basés sur la régression OPG en échantillons finis.

La classe des modèles à laquelle s'applique la DLR peut être écrite comme

$$f_t(y_t, \boldsymbol{\theta}) = \varepsilon_t, \quad t = 1, \dots, n, \quad \varepsilon_t \sim \text{NID}(0, 1), \quad (14.18)$$

où chaque  $f_t(\cdot)$  est une fonction lisse qui dépend de la variable aléatoire  $y_t$ , d'un vecteur de paramètres  $\boldsymbol{\theta}$  de dimension  $k$ , et (implicitement) de certaines variables exogènes et/ou prédéterminées. Comme la fonction  $f_t(\cdot)$  peut aussi dépendre des valeurs retardées de  $y_t$ , les modèles dynamiques sont permis. Ceci peut paraître à première vue être une classe de modèles plutôt restrictive, mais elle est en fait très générale. Par exemple, un modèle transformé des deux côtés, comme (14.05) peut, si les aléas sont supposés être  $\text{NID}(0, \sigma^2)$ , être écrit sous la forme de (14.18) en posant les définitions

$$f_t(y_t, \boldsymbol{\theta}) \equiv \frac{1}{\sigma} \left( \tau(y_t, \lambda) - \tau(x_t(\boldsymbol{\beta}), \lambda) \right) \quad \text{et} \quad \boldsymbol{\theta} \equiv [\boldsymbol{\beta} \vdash \lambda \vdash \sigma].$$

De la même manière, beaucoup d'autres modèles comportant des transformations de la variable dépendante peuvent être mis sous la forme (14.18). Il est même possible de mettre certains modèles multivariés sous cette forme; consulter Davidson et MacKinnon (1984a).

Pour un modèle de la classe à laquelle la DLR s'applique, la contribution de la  $i^{\text{ième}}$  observation à la fonction de logvraisemblance  $\ell(\mathbf{y}, \boldsymbol{\theta})$  est

$$\ell_t(y_t, \boldsymbol{\theta}) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} f_t^2(y_t, \boldsymbol{\theta}) + k_t(y_t, \boldsymbol{\theta}),$$

où

$$k_t(y_t, \boldsymbol{\theta}) \equiv \log \left| \frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial y_t} \right|$$



est un terme Jacobien. Maintenant établissons les définitions

$$F_{ti}(y_t, \boldsymbol{\theta}) \equiv \frac{\partial f_t(y_t, \boldsymbol{\theta})}{\partial \theta_i} \quad \text{et} \quad K_{ti}(y_t, \boldsymbol{\theta}) \equiv \frac{\partial k_t(y_t, \boldsymbol{\theta})}{\partial \theta_i}$$

et définissons  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$  et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  comme les matrices de dimension  $n \times k$  d'éléments type  $F_{ti}(y_t, \boldsymbol{\theta})$  et  $K_{ti}(y_t, \boldsymbol{\theta})$ . De façon similaire, soit  $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$  le vecteur de dimension  $n$  d'élément type  $f_t(y_t, \boldsymbol{\theta})$ . Il est facile de voir que le gradient de  $\ell(\mathbf{y}, \boldsymbol{\theta})$  est

$$\mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) = -\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta})\mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta})\boldsymbol{\iota}, \quad (14.19)$$

où  $\boldsymbol{\iota}$  désigne un vecteur de dimension  $n$  dont chaque élément est égale à 1.

Le résultat fondamental qui rend possible la DLR est que, pour cette classe de modèle, la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$  satisfait l'égalité

$$\mathcal{J}(\boldsymbol{\theta}) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} (\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta})\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta})\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})) \right) \quad (14.20)$$

et ainsi elle peut être estimée de façon convergente par

$$\frac{1}{n} (\mathbf{F}^\top(\mathbf{y}, \ddot{\boldsymbol{\theta}})\mathbf{F}(\mathbf{y}, \ddot{\boldsymbol{\theta}}) + \mathbf{K}^\top(\mathbf{y}, \ddot{\boldsymbol{\theta}})\mathbf{K}(\mathbf{y}, \ddot{\boldsymbol{\theta}})), \quad (14.21)$$

où  $\ddot{\boldsymbol{\theta}}$  est un estimateur quelconque convergent de  $\boldsymbol{\theta}$ . Nous nous sommes intéressés aux implications de (14.20) plutôt qu'à sa provenance. La dérivation fait appel à certaines propriétés plutôt spéciales de la distribution normale et peut être trouvée dans Davidson et MacKinnon (1984a).

La principale implication de (14.20) est qu'une certaine régression artificielle, que nous appelons la DLR, comporte toutes les propriétés que nous attendons obtenir d'une régression artificielle. La DLR peut être écrite comme

$$\begin{bmatrix} \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix} \mathbf{b} + \text{résidus}. \quad (14.22)$$

Cette régression artificielle comporte  $2n$  **observations artificielles**. La régres-sande est  $f_t(y_t, \boldsymbol{\theta})$  pour l'observation  $t$  et l'unité pour l'observation  $t+n$ , et les régresseurs correspondant au  $\boldsymbol{\theta}$  sont  $-\mathbf{F}_t(\mathbf{y}, \boldsymbol{\theta})$  pour l'observation  $t$  et  $\mathbf{K}_t(\mathbf{y}, \boldsymbol{\theta})$  pour l'observation  $t+n$ , où  $\mathbf{F}_t$  et  $\mathbf{K}_t$  désignent respectivement, les  $t^{\text{ièmes}}$  lignes de  $\mathbf{F}$  et de  $\mathbf{K}$ . De façon intuitive, la raison pour laquelle nous avons besoin ici d'une régression à longueur double est que chaque observation d'origine réalise deux contributions à la fonction de logvraisemblance : un terme de somme-des-carrés  $-\frac{1}{2}f_t^2$  et un terme Jacobien  $k_t$ . Nous savons comme résultat, qu'à la fois le gradient et la matrice d'information comprennent chacun deux parties, et la manière de tenir compte des deux à la fois consiste à incorporer deux observations artificielles dans la régression artificielle pour chaque observation d'origine.

Pourquoi (14.22) constitue-t-elle une régression artificielle valide? Comme nous l'avons noté lorsque nous discutons de la régression OPG dans la Section 13.7, il existe deux conditions principales qu'une régression artificielle doit satisfaire. Il est utile d'énoncer clairement ces conditions de manière quelque peu plus formelle ici.<sup>4</sup> Désignons  $\mathbf{r}(\mathbf{y}, \boldsymbol{\theta})$  la régressande pour une quelconque régression artificielle et  $\mathbf{R}(\mathbf{y}, \boldsymbol{\theta})$  la matrice des régresseurs. Soit  $n^*$  le nombre de lignes à la fois de  $\mathbf{r}(\mathbf{y}, \boldsymbol{\theta})$  et de  $\mathbf{R}(\mathbf{y}, \boldsymbol{\theta})$ , qui sera généralement soit  $n$ , soit un multiple entier de  $n$ . La régression de  $\mathbf{r}(\mathbf{y}, \boldsymbol{\theta})$  sur  $\mathbf{R}(\mathbf{y}, \boldsymbol{\theta})$  aura les propriétés d'une régression artificielle si

$$\mathbf{R}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}) = \rho(\boldsymbol{\theta}) \mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) \quad \text{et} \quad (14.23)$$

$$\text{plim}_{\boldsymbol{\theta}} \left( \frac{1}{n^*} \mathbf{R}^\top(\mathbf{y}, \check{\boldsymbol{\theta}}) \mathbf{R}(\mathbf{y}, \check{\boldsymbol{\theta}}) \right) = \rho(\boldsymbol{\theta}) \mathcal{J}(\boldsymbol{\theta}), \quad (14.24)$$

où  $\check{\boldsymbol{\theta}}$  désigne un estimateur convergent quelconque de  $\boldsymbol{\theta}$ . La notation  $\text{plim}_{\boldsymbol{\theta}}$  indique, comme d'habitude, que la limite en probabilité est donnée sous le DGP caractérisé par le vecteur de paramètres  $\boldsymbol{\theta}$ , et  $\rho(\boldsymbol{\theta})$  est un scalaire défini comme

$$\rho(\boldsymbol{\theta}) \equiv \text{plim}_{\boldsymbol{\theta}} \left( \frac{1}{n^*} \mathbf{r}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}) \right).$$

Parce que  $\rho(\boldsymbol{\theta})$  est égal à l'unité pour la régression OPG et pour la DLR, ces deux régressions artificielles satisfont des conditions plus simples

$$\mathbf{R}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{r}(\mathbf{y}, \boldsymbol{\theta}) = \mathbf{g}(\mathbf{y}, \boldsymbol{\theta}) \quad \text{et} \quad (14.25)$$

$$\text{plim}_{\boldsymbol{\theta}} \left( \frac{1}{n^*} \mathbf{R}^\top(\mathbf{y}, \check{\boldsymbol{\theta}}) \mathbf{R}(\mathbf{y}, \check{\boldsymbol{\theta}}) \right) = \mathcal{J}(\boldsymbol{\theta}), \quad (14.26)$$

aussi bien que les conditions d'origines (14.23) et (14.24). Cependant, ces conditions plus simples ne sont pas satisfaites par la GNR et sont donc de toute évidence trop simples en général.

Maintenant il est facile de voir que la DLR (14.21) satisfait les conditions (14.25) et (14.26). Pour la première de celles-ci, un simple calcul montre que

$$\begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) \\ \boldsymbol{\iota} \end{bmatrix} = -\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{f}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta}) \boldsymbol{\iota},$$

qui par (14.19) est égal au gradient  $\mathbf{g}(\mathbf{y}, \boldsymbol{\theta})$ . Pour la seconde, nous voyons que

$$\begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix}^\top \begin{bmatrix} -\mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \\ \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}) \end{bmatrix} = \mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) + \mathbf{K}^\top(\mathbf{y}, \boldsymbol{\theta}) \mathbf{K}(\mathbf{y}, \boldsymbol{\theta}).$$

<sup>4</sup> Pour un traitement plus complet sur ce sujet, consulter Davidson et MacKinnon (1990).

Le membre de droite est juste l'expression qui apparaît dans le résultat fondamental (14.20). En conséquence, il est clair que la DLR doit satisfaire (14.26). Toute cette discussion suppose, naturellement, que les matrices  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$  et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  satisfont les conditions de régularité appropriées, qui peuvent ne pas être toujours facilement vérifiables en la réalité; consulter Davidson et MacKinnon (1984a).

La DLR peut être utilisée de toutes les manières que la GNR et la régression OPG peuvent être utilisées. En particulier, elle peut être utilisée pour

- (i) vérifier que les conditions du premier ordre pour un maximum d'une fonction de logvraisemblance sont satisfaites de façon suffisamment exactes,
- (ii) calculer les matrices de covariance estimées,
- (iii) calculer les statistiques de test,
- (iv) calculer les estimations efficaces en une étape, et
- (v) comme partie clé des procédures d'estimation ML.

L'utilisation de (i) a été discutée dans le contexte de la GNR dans la Section 6.1; celle de (ii) a été abordée dans les Sections 6.2, 10.4, et 13.7; l'emploi de (iii) a été beaucoup usité tout au long du livre, et notamment au début du Chapitre 6; et les utilisations de (iv) et (v) ont été discutées, dans le contexte de la GNR, dans les Sections 6.6 et 6.8. Tout ce qui a été dit, ou presque, concernant les utilisations de la GNR et de la régression OPG s'applique également aussi bien à la DLR et ne sera pas, par conséquent, répété ici.

De nombreuses statistiques de test différentes peuvent être programmées en utilisant la même régression artificielle à longueur double. Dans sa forme score, la statistique LM est

$$\tilde{\mathbf{g}}^\top (n\tilde{\mathbf{J}})^{-1} \tilde{\mathbf{g}}, \quad (14.27)$$

où  $\tilde{\mathbf{g}} \equiv \mathbf{g}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$  est le gradient évalué en un ensemble d'estimations contraintes  $\tilde{\boldsymbol{\theta}}$ . Si nous lançons la DLR (14.22) avec les quantités  $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ ,  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$ , et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  évaluées en  $\tilde{\mathbf{f}} \equiv \mathbf{f}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ ,  $\tilde{\mathbf{F}} \equiv \mathbf{F}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ , et  $\tilde{\mathbf{K}} \equiv \mathbf{K}(\mathbf{y}, \tilde{\boldsymbol{\theta}})$ , la somme expliquée des carrés sera

$$(-\tilde{\mathbf{f}}^\top \tilde{\mathbf{F}} + \boldsymbol{\iota}^\top \tilde{\mathbf{K}})(\tilde{\mathbf{F}}^\top \tilde{\mathbf{F}} + \tilde{\mathbf{K}}^\top \tilde{\mathbf{K}})^{-1}(-\tilde{\mathbf{F}}^\top \tilde{\mathbf{f}} + \tilde{\mathbf{K}}^\top \boldsymbol{\iota}). \quad (14.28)$$

Ceci a clairement la même forme que la statistique LM (14.27). A partir de (14.19), nous voyons que  $\tilde{\mathbf{g}} = -\tilde{\mathbf{F}}^\top \tilde{\mathbf{f}} + \tilde{\mathbf{K}}^\top \boldsymbol{\iota}$ . A partir de (14.20), nous voyons que  $\mathcal{J}(\boldsymbol{\theta})$  est estimée de façon convergente par  $n^{-1}(\tilde{\mathbf{F}}^\top \tilde{\mathbf{F}} + \tilde{\mathbf{K}}^\top \tilde{\mathbf{K}})$  quand les restrictions sont vraies. Ainsi, la somme expliquée des carrés à partir de la DLR, expression (14.28), fournira une statistique de test valide asymptotiquement. Comme d'habitude, les statistiques pseudo- $F$  et pseudo- $t$  seront également valides.

L'expression générale d'une DLR, (14.22), est d'une simplicité trompeuse. Il peut être alors intéressant de voir ce qui se passe si nous utilisons une DLR

dans un cas simple que nous savons déjà traiter. Considérons un modèle de régression non linéaire univarié

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Lorsqu'il est écrit sous la forme de (14.18), ce modèle devient

$$f_t(y_t, \boldsymbol{\theta}) \equiv \frac{1}{\sigma}(y_t - x_t(\boldsymbol{\beta})) = \varepsilon_t. \quad (14.29)$$

Si  $\boldsymbol{\beta}$  est un vecteur de dimension  $k$ ,  $\boldsymbol{\theta}$  sera un vecteur de dimension  $(k + 1)$ . Considérons maintenant la façon dont nous pourrions tester les restrictions sur  $\boldsymbol{\beta}$  en utilisant une DLR. La nature et le nombre des restrictions sont non pertinents pour notre propos; pour faire simple, nous pouvons supposer qu'on a  $r \leq k$  restrictions de nullité. Les quantités désignées par  $\sim$  sont évaluées en des estimations ML (par exemple, NLS) soumises à ces restrictions.

En calculant  $\mathbf{f}(\mathbf{y}, \boldsymbol{\theta})$ ,  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$ , et  $\mathbf{K}(\mathbf{y}, \boldsymbol{\theta})$  pour le modèle (14.29), en les évaluant aux estimations contraintes  $\tilde{\boldsymbol{\theta}}$ , et en substituant les résultats dans (14.22), cela fournit la DLR

$$\begin{bmatrix} \tilde{\boldsymbol{\varepsilon}} \\ \boldsymbol{\iota} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}}/\tilde{\sigma} & \tilde{\boldsymbol{\varepsilon}}/\tilde{\sigma} \\ \mathbf{0} & -\boldsymbol{\iota}/\tilde{\sigma} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ s \end{bmatrix} + \text{résidus}. \quad (14.30)$$

Ici  $\tilde{\boldsymbol{\varepsilon}} \equiv \tilde{\mathbf{u}}/\tilde{\sigma}$  désigne un vecteur de dimension  $n$  de résidus normalisés et  $\tilde{\mathbf{X}}$  désigne une matrice de dimension  $n \times k$  d'élément type  $\partial x_t(\boldsymbol{\beta})/\partial \beta_i$ , évalué en  $\tilde{\boldsymbol{\beta}}$ . Les  $k$  premiers régresseurs dans (14.30) correspondent aux éléments de  $\boldsymbol{\beta}$ , tandis que le dernier correspond à  $\sigma$ ; Ils ont respectivement les coefficients  $\mathbf{b}$  et  $s$ . Il est évident que le dernier régresseur est orthogonal à la régressande. Il est aussi orthogonal à tous les régresseurs qui correspondent aux éléments de  $\boldsymbol{\beta}$  qui ont été estimés sans restriction (par les conditions du premier ordre) et, sous l'hypothèse nulle, il devrait être non corrélé avec les régresseurs restant. Ainsi il est asymptotiquement valide de laisser tomber ce dernier régresseur. Mais quand il est tombé, la seconde moitié de la DLR devient non pertinente, puisque dans la seconde moitié, tous les régresseurs qui restent sont nuls. Si les facteurs de  $1/\tilde{\sigma}$  sont ignorés, il nous reste la régression artificielle

$$\tilde{\mathbf{u}} = \tilde{\mathbf{X}}\mathbf{b} + \text{résidus}, \quad (14.31)$$

qui est simplement la régression de Gauss-Newton. Comme la régressande n'est pas divisée par  $\tilde{\sigma}$ , il est maintenant nécessaire de diviser la somme expliquée des carrés de (14.31) par une estimation de  $\sigma^2$  quand nous calculons la statistique de test.

Le fait que la DLR est équivalente à la GNR quand cette dernière est valide fait sens. Supposons que  $\text{ESS}_{\text{DLR}}$  désigne la somme expliquée des carrés provenant de (14.30) et que  $\text{ESS}_{\text{GN}}$  désigne la somme expliquée des carrés provenant de la GNR modifiée obtenue à partir de (14.31) en remplaçant  $\tilde{\mathbf{u}}$

par  $\tilde{\varepsilon}$ . Il peut être montré que ces deux statistiques de test sont toutes deux des fonctions de la même variable aléatoire. Cependant, elles *ne seront pas* numériquement identiques, la relation exacte entre elles étant

$$\text{ESS}_{\text{DLR}} = \frac{\text{ESS}_{\text{GN}}}{1 - \text{ESS}_{\text{GN}}/(2n)}.$$

Comme  $\text{ESS}_{\text{DLR}}$  sera toujours plus grande que  $\text{ESS}_{\text{GN}}$ , la DLR sera toujours quelque chose de plus enclin à rejeter l'hypothèse nulle que la régression de Gauss-Newton. La différence entre elles sera habituellement très petite, à moins que  $n$  ne soit très petit ou que  $\text{ESS}_{\text{GN}}$  ne soit très grande. Si, à la place de la somme expliquée des carrés, les statistiques  $t$  ou  $F$  sont utilisées, il peut être montré que la DLR et les régressions de Gauss-Newton fournissent des résultats numériquement identiques, sauf pour des corrections légèrement différentes pour des degrés de liberté.

Il est inutile naturellement, d'utiliser une DLR quand une GNR s'applique, c'est-à-dire quand à la fois l'hypothèse nulle et l'hypothèse alternative sont des modèles de régression. Mais quand la variable dépendante est soumise à une transformation non linéaire qui dépend de paramètres inconnus, la GNR n'est pas applicable. Dans la prochaine section, nous montrons comment la DLR peut être utilisée avec les modèles de Box-Cox et d'autres modèles qui comportent des transformations de la variable dépendante.

## 14.5 LA DLR ET LES VARIABLES TRANSFORMÉES

Il est facile de déterminer la forme spécifique que prend la DLR pour chacun des modèles (14.04), (14.05), et (14.06) pour n'importe quelle transformation spécifiée  $\tau(y_t, \lambda)$ . Considérons (14.04) tout d'abord. Nous pouvons écrire

$$f_t(y_t, \beta, \lambda, \sigma) \equiv \frac{1}{\sigma}(\tau(y_t, \lambda) - x_t(\beta)).$$

A partir de (14.22), nous voyons que la régressande pour la DLR est

$$\mathbf{r}(\theta) = \begin{bmatrix} f_t(y_t, \beta, \lambda, \sigma) \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma}(\tau(y_t, \lambda) - x_t(\beta)) \\ 1 \end{bmatrix},$$

où les quantités supérieures et inférieures à l'intérieur des grands crochets désignent, respectivement, le  $i^{\text{ième}}$  élément et le  $(t+n)^{\text{ième}}$  de la régressande. Nous utiliserons cette notation de façon extensive quand nous discuterons des DLR.

Pour les trois modèles — (14.04), (14.05), et (14.06) — le terme Jacobien pour la  $t^{\text{ième}}$  observation est

$$k_t \equiv \log \left( \frac{\partial f_t(y_t, \beta, \lambda, \sigma)}{\partial y_t} \right) = \log(\tau_y(y_t, \lambda)) - \log \sigma,$$

où  $\tau_y(y_t, \lambda)$  désigne  $\partial\tau(y_t, \lambda)/\partial y_t$ . Ainsi, la matrice des régresseurs pour la DLR qui correspond à (14.04) est

$$\mathbf{R}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma} \mathbf{X}_t(\boldsymbol{\beta}) & -\frac{1}{\sigma} \tau_\lambda(y_t, \lambda) & \frac{\tau(y_t, \lambda) - x_t(\boldsymbol{\beta})}{\sigma^2} \\ \mathbf{0} & \frac{\tau_{y\lambda}(y_t, \lambda)}{\tau_y(y_t, \lambda)} & -\frac{1}{\sigma} \end{bmatrix}, \quad (14.32)$$

où  $\tau_\lambda(y_t, \lambda)$  désigne  $\partial\tau(y_t, \lambda)/\partial\lambda$ , et  $\tau_{y\lambda}(y_t, \lambda)$  désigne  $\partial\tau_y(y_t, \lambda)/\partial\lambda$ . Les deux quantités dans la première colonne de (14.32) désignent les  $t^{\text{ième}}$  et  $(t+n)^{\text{ième}}$  lignes des  $k$  colonnes de la matrice des régresseurs qui correspond à  $\boldsymbol{\beta}$ . De façon similaire, les deux quantités dans chacune des deuxième et troisième colonnes désignent les éléments de la matrice des régresseurs qui correspondent à  $\lambda$  et  $\sigma$ , respectivement.

Quand la transformation  $\tau$  est la transformée de Box-Cox,

$$\tau_\lambda(y, \lambda) = \frac{\lambda y^\lambda \log y - y^\lambda + 1}{\lambda^2} \quad \text{et}$$

$$\frac{\tau_{y\lambda}(y, \lambda)}{\tau_y(y, \lambda)} = \frac{y^{\lambda-1} \log(y)}{y^{\lambda-1}} = \log(y).$$

Par conséquent, la DLR pour le modèle de Box-Cox simple (14.04) avec  $\tau(y_t, \lambda)$  donnée par la transformée de Box-Cox, est

$$\begin{bmatrix} \frac{1}{\sigma} u_t(y_t, \boldsymbol{\beta}, \lambda) \\ 1 \end{bmatrix} \quad (14.33)$$

$$= \begin{bmatrix} \frac{1}{\sigma} \mathbf{X}_t(\boldsymbol{\beta}) & \frac{-(\lambda y_t^\lambda \log y_t - y_t^\lambda + 1)}{\sigma \lambda^2} & \frac{u_t(y_t, \boldsymbol{\beta}, \lambda)}{\sigma^2} \\ \mathbf{0} & \log y_t & -\frac{1}{\sigma} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ a \\ s \end{bmatrix} + \text{résidus},$$

où  $\mathbf{b}$  est un vecteur de dimension  $k$  des coefficients qui correspondent au  $\boldsymbol{\beta}$ ,  $a$  et  $s$  sont des coefficients scalaires qui correspondent à  $\lambda$  et à  $\sigma$ , et

$$u_t(y_t, \boldsymbol{\beta}, \lambda) \equiv B(y_t, \lambda) - x_t(\boldsymbol{\beta}).$$

Si la DLR (14.33) est évaluée en des estimations ML non contraintes  $\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma})$ , tous les coefficients estimés seront nuls. Puisque les conditions du premier ordre pour  $\sigma$  impliquent que

$$\hat{\sigma} = \left( \frac{1}{n} \sum_{t=1}^n \hat{u}_t^2 \right)^{1/2},$$

la somme totale des carrés à partir de la régression artificielle sera  $2n$ . Ainsi, l'estimation de la matrice de covariance OLS sera simplement  $(2n/(2n - k - 2))(\hat{\mathbf{R}}^\top \hat{\mathbf{R}})^{-1}$ , où  $\hat{\mathbf{R}}$  désigne la matrice des régresseurs qui apparaît dans (14.33), évaluée aux estimations ML. D'après le résultat fondamental (14.20), cette matrice de covariance OLS fournit une estimation valide de la matrice de covariance asymptotique de l'estimateur ML  $\hat{\boldsymbol{\theta}}$ .

Il est clair à partir de (14.33) que cette matrice de covariance asymptotique n'est pas bloc diagonale entre  $\boldsymbol{\beta}$  et les autres paramètres. En formant la matrice  $\mathbf{R}^\top \mathbf{R}$ , en divisant par  $n$ , et en prenant les limites en probabilité, nous voyons que le bloc  $(\boldsymbol{\beta}, \boldsymbol{\beta})$  de la matrice d'information  $\mathcal{J}(\boldsymbol{\theta})$  est simplement

$$\sigma^{-2} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top(\boldsymbol{\beta}) \mathbf{X}(\boldsymbol{\beta}) \right), \quad (14.34)$$

comme cela serait le cas s'il s'agissait d'un modèle de régression non linéaire. L'élément  $(\sigma, \sigma)$  est simplement  $2/\sigma^2$ , qui est encore ce qu'il serait s'il y avait un modèle de régression non linéaire. Mais  $\mathcal{J}(\boldsymbol{\theta})$  contient aussi un élément  $(\lambda, \lambda)$  un élément  $(\lambda, \sigma)$ , une ligne et une colonne  $(\boldsymbol{\beta}, \lambda)$ , chacun d'entre eux étant clairement non nul. Par exemple, l'élément qui correspond à  $\beta_i$  et  $\lambda$  est

$$- \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n\sigma^2\lambda^2} \sum_{t=1}^n X_{ti}(\boldsymbol{\beta}) (\lambda y_t^\lambda \log y_t - y_t^\lambda + 1) \right).$$

Les éléments  $(\lambda, \lambda)$  et  $(\lambda, \sigma)$  peuvent aussi être obtenus d'une manière directe et ils sont manifestement différents de zéro.

Comme  $\mathcal{J}(\boldsymbol{\theta})$  n'est pas bloc diagonale entre  $\boldsymbol{\beta}$  et les deux autres paramètres, le bloc  $(\boldsymbol{\beta}, \boldsymbol{\beta})$  de son inverse ne sera pas égal à l'inverse de (14.34). Ainsi, comme nous l'avons précisé dans la Section 14.2, il est incorrect d'établir des inférences en utilisant la matrice de covariance estimée NLS pour  $\boldsymbol{\beta}$  conditionnelle à  $\lambda$ . De façon similaire, comme l'élément  $(\lambda, \sigma)$  de  $\mathcal{J}(\boldsymbol{\theta})$  est non nul, nous pouvons trouver l'inverse du bloc de dimension  $(k+1) \times (k+1)$  de la matrice d'information qui correspond à  $\boldsymbol{\beta}$  et  $\lambda$  conjointement sans inverser complètement la matrice d'information. La matrice de covariance estimée obtenue en employant une application NLS en donnant des estimations ML sera donc incorrecte.

Il devrait être clair que ce dont nous venons de parler concernant le modèle de Box-Cox simple s'applique également au modèle transformé des deux côtés et au modèle conventionnel, puisque le Jacobien de la transformation est le même pour tous ces modèles. Il est facile d'établir les DLR pour les deux autres modèles. Dans les deux cas, la régressande a la même forme que la régressande de (14.33), sauf pour le modèle transformé des deux côtés

$$u_t(y_t, \boldsymbol{\beta}, \lambda) \equiv B(y_t, \lambda) - B(x_t(\boldsymbol{\beta}), \lambda)$$

et pour le modèle de Box-Cox conventionnel

$$u_t(y_t, \boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda) \equiv B(y_t, \lambda) - \sum_{i=1}^k \beta_i B(X_{ti}, \lambda) - \sum_{j=1}^l \gamma_j Z_{tj}.$$

Le régresseur qui correspond à  $\sigma$  a aussi la même forme que celle qui apparaît dans (14.33).

Pour le modèle transformé des deux côtés, le régresseur qui correspond à  $\beta_i$  est

$$\begin{bmatrix} \frac{1}{\sigma} (x_t(\boldsymbol{\beta}))^{\lambda-1} X_{ti}(\boldsymbol{\beta}) \\ \mathbf{0} \end{bmatrix},$$

et le régresseur qui correspond à  $\lambda$  est

$$\begin{bmatrix} \frac{1}{\sigma \lambda^2} \left( (\lambda (x_t(\boldsymbol{\beta}))^\lambda \log(x_t(\boldsymbol{\beta})) - (x_t(\boldsymbol{\beta}))^\lambda + 1) - (\lambda y_t^\lambda \log y_t - y_t^\lambda + 1) \right) \\ \log y_t \end{bmatrix}.$$

Pour le modèle de Box-Cox conventionnel, les régresseurs qui correspondent à  $\beta_i$  et  $\gamma_j$ , respectivement, sont

$$\begin{bmatrix} \frac{1}{\sigma} B(X_{ti}, \lambda) \\ 0 \end{bmatrix} \quad \text{et} \quad \begin{bmatrix} \frac{1}{\sigma} Z_{tj} \\ 0 \end{bmatrix}, \quad (14.35)$$

et le régresseur qui correspond à  $\lambda$  est

$$\begin{bmatrix} \frac{1}{\sigma \lambda^2} \left( \sum_{i=1}^k \beta_i (\lambda X_{ti}^\lambda \log X_{ti} - X_{ti}^\lambda + 1) - (\lambda y_t^\lambda \log y_t - y_t^\lambda + 1) \right) \\ \log y_t \end{bmatrix}. \quad (14.36)$$

Nous avons maintenant obtenu des DLR pour les trois types les plus communs des modèles de Box-Cox. Des DLR pour d'autres types de modèles qui comprennent des transformations de la variable dépendante peuvent être dérivées de façon similaire. Toutes ces DLR peuvent être utilisées comme une partie clé des algorithmes pour estimer les modèles auxquels ils s'appliquent, exactement de la même manière que les GNR peuvent être utilisées comme une partie des algorithmes pour estimer les modèles de régression non linéaire; consulter la Section 6.8. Etant donnée une quelconque valeur de  $\lambda$  (très probablement 0 ou 1), il est facile d'obtenir des estimations initiales des paramètres restant du modèle par OLS ou NLS. Ceci fournit alors un ensemble complet d'estimations paramétriques, disons  $\boldsymbol{\theta}^{(1)}$ , auquel la DLR peut être évaluée au début. Les estimations des coefficients à partir de la DLR, disons  $\boldsymbol{t}^{(1)}$ , peuvent alors être utilisées pour déterminer la direction dans laquelle mettre à jour les estimations paramétriques, et le processus entier peut être répété autant de fois que nécessaire jusqu'à ce qu'une règle d'arrêt soit satisfaite.



La règle d'actualisation de l'algorithme de maximisation a la forme

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \alpha^{(j)} \mathbf{t}^{(j)}. \quad (14.37)$$

Ici  $\boldsymbol{\theta}^{(j)}$  et  $\boldsymbol{\theta}^{(j+1)}$  désignent les vecteurs des estimations sur les  $j^{\text{ième}}$  et  $(j+1)^{\text{ième}}$  itérations de l'algorithme de maximisation,  $\mathbf{t}^{(j)}$  désigne le vecteur des coefficients estimés à partir de la DLR, et  $\alpha^{(j)}$  désigne la longueur de pas, qui peut être choisie de diverses manières par l'algorithme. Cette règle d'actualisation ressemble à celle de la régression de Gauss-Newton discutées dans la Section 6.8, et fonctionne pour exactement la même raison. Un algorithme basé sur la méthode de Newton (avec une longueur de pas variable  $\alpha$ ) utiliserait la règle d'actualisation

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \alpha^{(j)} (\mathbf{H}(\boldsymbol{\theta}^{(j)}))^{-1} \mathbf{g}(\boldsymbol{\theta}^{(j)}). \quad (14.38)$$

La DLR à pas  $j$  fournit le vecteur de coefficients

$$\mathbf{t}^{(j)} = (\mathbf{R}^\top(\boldsymbol{\theta}^{(j)}) \mathbf{R}(\boldsymbol{\theta}^{(j)}))^{-1} \mathbf{R}^\top(\boldsymbol{\theta}^{(j)}) \mathbf{r}(\boldsymbol{\theta}^{(j)}).$$

D'après la propriété de toutes les régressions artificielles,  $\mathbf{t}^{(j)}$  est asymptotiquement égale à moins l'inverse du Hessien fois le gradient. Par conséquent, il est sensé remplacer  $-(\mathbf{H}(\boldsymbol{\theta}^{(j)}))^{-1} \mathbf{g}(\boldsymbol{\theta}^{(j)})$  dans (14.38) par  $\mathbf{t}^{(j)}$ . Ce qui donne (14.37), qui est la règle d'actualisation basée sur la DLR. La règle d'arrêt devrait normalement être fondée sur une certaine mesure du pouvoir explicatif de la DLR, cela a été discuté dans la Section 6.8.

La DLR peut, naturellement, être utilisée pour la mise en œuvre de tests d'hypothèse de n'importe lequel des modèles dont nous avons discuté. Puisque pour ces modèles la somme des carrés de la régressande est toujours  $2n$ , la quantité  $2n - \text{SSR}$  égalera toujours la somme expliquée des carrés, et elle fournit une statistique de test valide asymptotiquement qui est très facile à calculer. Comme d'habitude, les statistiques pseudo- $F$  et pseudo- $t$  basées sur la régression artificielle sont également valides asymptotiquement. Nous n'élaborerons pas ces sujets ici, puisque qu'il n'y a rien de nouveau pour en discuter; un cas spécifique sera discuté dans la prochaine section.

Il est peut être bon d'interjeter une petite mise en garde sur ce point. Si la régressande ou certains régresseurs dans une DLR qui est utilisée pour tester les hypothèses est construite de façon incorrecte, il est possible, et en effet très probable, que la régression fournira une statistique de test calculée grande et dénuée de sens. Contrôler la plupart des calculs en lançant tout d'abord la DLR *sans* ces régresseurs constitue alors une très bonne idée; ces derniers correspondent aux paramètres qui ont été testés. Cette régression, tout comme les régressions artificielles utilisées pour calculer les matrices de covariance, devrait n'avoir aucun pouvoir explicatif si tout a été construit correctement. Malheureusement, nous ne pouvons pas vérifier les régresseurs

de test de cette manière, et une erreur dans leur construction peut facilement conduire à des résultats incohérents. Par exemple, si nous rajoutons par inadvertance un terme constant à la DLR, il aura presque certainement un pouvoir explicatif substantiel sur la régressande, parce que la seconde moitié de cette dernière est simplement un vecteur composé de 1.

## 14.6 TEST DE MODÈLES LINÉAIRE ET LOGLINÉAIRE

Dans de nombreuses applications, la variable dépendante est toujours positive. Les économètres doivent alors décider si un modèle de régression devrait tenter d'expliquer l'espérance de la variable originale ou de son logarithme. Les deux types de modèles sont souvent plausibles *a priori*. Dans cette section, nous discutons de techniques de sélection entre modèles dans lesquels la régressande est le niveau ou le logarithme de la variable dépendante, et de techniques de test de leur spécification. Les tests basés sur la DLR s'avèrent très utiles pour ce propos.

Supposons initialement que les deux modèles sont linéaires en leurs paramètres. Ainsi, les deux modèles qui se font concurrence sont

$$y_t = \sum_{i=1}^k \beta_i X_{ti} + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad \text{et} \quad (14.39)$$

$$\log y_t = \sum_{i=1}^k \beta_i \log X_{ti} + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad (14.40)$$

où la notation, non pas par coïncidence, est la même que pour le modèle de Box-Cox conventionnel. Après l'estimation des deux modèles, il peut être possible de conclure que l'un d'eux devrait être rejeté simplement en comparant les valeurs de leurs fonctions de logvraisemblance, comme cela a été discuté dans la Section 14.3. Cependant, une telle procédure ne peut rien nous dire concernant la validité du modèle qui s'ajuste le mieux. Si les deux modèles sont raisonnables, il est important de les tester tous les deux avant de tenter d'en accepter un autre.

Il existe de nombreuses manières de tester la spécification des modèles de régression linéaire et non linéaire comme (14.39) et (14.40). Les tests les plus communément usités sont basés sur le fait que ceux-ci sont tous les deux des cas spéciaux du modèle de Box-Cox conventionnel

$$B(y_t, \lambda) = \sum_{i=1}^k \beta_i B(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \quad (14.41)$$

De façon conceptuelle, la manière la plus simple de tester (14.39) et (14.40) contre (14.41) consiste à estimer les trois modèles et d'employer un test LR,

comme cela fut suggéré à l'origine par Box et Cox (1964) dans contexte du modèle de Box-Cox simple. Cependant, comme l'estimation de (14.41) peut demander un certain effort, il peut être plus intéressant d'utiliser un test LM à sa place.

Plusieurs manières d'implémenter ce test LM sont valables. Nous ne mentionnerons seulement que celles basées sur les régressions artificielles, puisque celles-ci sont les plus simples à calculer, et, si un test LM n'est pas facile à calculer, il n'a aucun avantage sur le test correspondant LR. Il est évidemment possible de construire des tests LM de (14.39) et (14.40) contre (14.41) en utilisant soit la régression OPG soit la DLR. Les premiers tests tirent leur origine de Godfrey et Wickens (1981) et les derniers de Davidson et MacKinnon (1985c). Les derniers auteurs ont fourni des évidences Monte Carlo que les tests basés sur la DLR sont beaucoup plus performants en échantillons finis que ceux basés sur la régression OPG, une avancée confirmée plus tard par Godfrey, McAleer, et McKenzie (1988).

Il est intéressant de discuter à quoi ressemble la DLR pour tester les régressions linéaire et loglinéaire. Quand nous testons le modèle linéaire (14.39), l'hypothèse nulle est que  $\lambda = 1$ . Dans ce cas, la régressande de la DLR comporte le  $i^{\text{ième}}$  élément  $\hat{u}_t/\hat{\sigma}$  et le  $(t+n)^{\text{ième}}$  élément 1, où  $\hat{u}_t$  désigne le  $t^{\text{ième}}$  résidu issu du modèle linéaire et  $\hat{\sigma}$  désigne l'estimation ML de  $\sigma$ . Le  $t^{\text{ième}}$  et le  $(t+n)^{\text{ième}}$  éléments des régresseurs sont alors

pour  $\beta_i$  :  $X_{ti} - 1$  et 0;

pour  $\gamma_j$  :  $Z_{tj}$  et 0;

pour  $\sigma$  :  $\hat{u}_t/\hat{\sigma}$  et  $-1$ ;

pour  $\lambda$  :  $\sum_{i=1}^k \hat{\beta}_i (X_{ti} \log X_{ti} - X_{ti} + 1) - (y_t \log y_t - y_t + 1)$  et  $\hat{\sigma} \log y_t$ .

Ces régresseurs ne correspondent pas à ceux auxquels l'on pourrait s'attendre à avoir à partir de (14.33), (14.35), et (14.36), parce qu'ils ont tous été multipliés par  $\hat{\sigma}$ , quelque chose qui est sans effet parce qu'il ne change pas le sous-espace engendré par les colonnes de la matrice de régresseurs. Pour la même raison, si un des  $Z_{tj}$  est un terme constant, comme cela sera typiquement le cas, il n'est pas nécessaire de soustraire 1 des  $X_{ti}$ .

Quand nous testons le modèle loglinéaire (14.40), l'hypothèse nulle est que  $\lambda = 0$ . Dans ce cas, la régressande de la DLR comprend le  $i^{\text{ième}}$  élément  $\tilde{u}_t/\tilde{\sigma}$  et le  $(t+n)^{\text{ième}}$  élément 1, où  $\tilde{u}_t$  désigne le  $i^{\text{ième}}$  résidu provenant du modèle loglinéaire et  $\tilde{\sigma}$  désigne l'estimation ML de  $\sigma$ . Le  $i^{\text{ième}}$  et le  $(t+n)^{\text{ième}}$

éléments des régresseurs sont alors

pour  $\beta_i$  :  $\log X_{ti}$  et 0;

pour  $\gamma_j$  :  $Z_{tj}$  et 0;

pour  $\sigma$  :  $\tilde{u}_t/\tilde{\sigma}$  et  $-1$ ;

pour  $\lambda$  :  $\frac{1}{2} \sum_{i=1}^k \tilde{\beta}_i (\log X_{ti})^2 - \frac{1}{2} (\log y_t)^2$  et  $\tilde{\sigma} \log y_t$ .

Cette fois tous les régresseurs ont été multipliés par  $\tilde{\sigma}$ . Le régresseur pour  $\lambda$  avait trouvé son origine à l'aide de la Règle de l'Hôpital :

$$\lim_{\lambda \rightarrow 0} \left( \frac{\lambda x^\lambda \log x - x^\lambda + 1}{\lambda^2} \right) = \frac{1}{2} (\log x)^2.$$

Un test, qui est parfois confondu avec les tests LM tels que ceux dont nous venons juste de discuter, est un test proposé par Andrews (1971) et modifié par Godfrey et Wickens (1981) de telle sorte qu'il s'applique au modèle de Box-Cox conventionnel. L'idée est de prendre une approximation du premier ordre de (14.41) autour de  $\lambda = 0$  ou  $\lambda = 1$ , de réarranger les termes de telle sorte que seul  $\log y_t$  ou  $y_t$  apparaisse sur le membre de gauche, et alors remplacer  $y_t$  à chaque fois qu'il apparaît sur le membre de droite par les valeurs ajustées provenant de la régression sous test. Le résultat est quelque chose qui ressemble à la régression originale qui a été testée, avec l'addition d'un régresseur supplémentaire. Pour l'hypothèse nulle linéaire, ce régresseur supplémentaire est

$$\hat{y}_t \log \hat{y}_t - \hat{y}_t + 1 - \sum_{i=1}^k \hat{\beta}_i (X_{ti} \log X_{ti} - X_{ti} + 1),$$

et pour l'hypothèse loglinéaire il est

$$\frac{1}{2} \left( (\log \tilde{y}_t)^2 - \sum_{i=1}^k \tilde{\beta}_i (\log X_{ti})^2 \right),$$

où  $\hat{y}_t$  et  $\tilde{y}_t$  désignent les valeurs ajustées de  $y_t$  issues des modèles linéaire et loglinéaire respectivement. La statistique de test est simplement le  $t$  de Student sur le régresseur supplémentaire.

Le test de Andrews comporte une propriété plutôt remarquable. Si les  $X_{ti}$  et les  $Z_{tj}$  peuvent être traités comme exogènes, et si les aléas sont réellement normalement distribués, la statistique de test aura réellement la distribution de Student en échantillons finis. Ceci provient du fait que les régresseurs de test dépendent de  $y_t$  seulement au travers des estimations  $\hat{\beta}$  et  $\hat{\gamma}$  (ou  $\tilde{\beta}$  et  $\tilde{\gamma}$ ). L'argument est similaire à celui utilisé dans la Section 11.3 pour montrer que le test  $J_A$  est exact. Il s'ensuit à partir des mêmes résultats de Milliken et de Graybill (1970).

Cependant, le test d'Andrews n'est pas véritablement un test contre la même alternative que les tests LM. De façon implicite, il teste une direction de régression, contre une alternative qui est aussi un modèle de régression. Mais le modèle de Box-Cox (14.41) n'est pas un modèle de régression. Le test d'Andrews doit avoir par conséquent moins de puissance que les tests classiques des modèles linéaire et loglinéaire contre (14.41) quand le dernier a réellement généré les données. En utilisant des techniques similaires à celles discutées dans le Chapitre 12, il a été montré dans Davidson et MacKinnon (1985c) que, quand  $\sigma \rightarrow 0$ , le paramètre de non centralité pour le test d'Andrews s'approche de celui des tests classiques, tandis que, quand  $\sigma \rightarrow \infty$ , il s'approche de zéro. Ainsi, sauf quand  $\sigma$  est petit nous devrions nous attendre à ce que le test d'Andrews manque sérieusement de puissance, et les résultats Monte Carlo confirment cela. Cependant, un avantage possible du test d'Andrews devrait être noté. Contrairement aux tests LM dont nous avons discuté, il n'est pas sensible, asymptotiquement, aux caractéristiques de l'hypothèse de normalité, parce qu'il teste simplement une direction de régression.

Bien que les tests basés sur la transformée de Box-Cox soient très populaires, une seconde approche pour tester les modèles linéaire et loglinéaire mérite aussi d'être mentionnée. Elle traite les deux modèles comme des hypothèses non emboîtées, de la même manière que cela a été fait pour les tests discutés dans la Section 11.3. Cette approche non emboîtée permet de traiter des types plus généraux de modèle que l'approche basée sur la transformée de Box-Cox, parce qu'il n'est pas nécessaire que les deux modèles aient le même nombre de paramètres, ou qu'ils se ressemblent d'une quelconque manière, et il n'est pas nécessaire non plus qu'ils soient linéaires par rapport aux variables ou aux paramètres. Nous pouvons écrire les deux modèles en compétition comme

$$H_1: y_t = x_t(\beta) + u_{1t}, \quad u_{1t} \sim \text{NID}(0, \sigma_1^2), \quad \text{et} \quad (14.42)$$

$$H_2: \log y_t = z_t(\gamma) + u_{2t}, \quad u_{2t} \sim \text{NID}(0, \sigma_2^2). \quad (14.43)$$

Ici, la notation est similaire à celle qui a été utilisée pour le test d'hypothèses non emboîtées dans la Section 11.3 et devrait s'expliquer d'elle-même. Notons que l'hypothèse selon laquelle les aléas suivent une loi normale, dont nous n'avons pas besoin dans notre discussion précédente, est ici nécessaire.

Il existe deux manières évidentes de construire des tests non emboîtés pour les modèles comme (14.42) et (14.43). La première est de tenter d'implémenter les idées de Cox (1961, 1962), comme dans Aneuryn-Evans et Deaton (1980). Malheureusement, elle s'avère plutôt difficile. La seconde approche, beaucoup plus facile, consiste à les baser sur une certaine sorte d'emboîtement artificiel. Considérons (quelque peu arbitrairement) le modèle composite artificielle

$$H_C: (1 - \alpha) \left( \frac{y_t - x_t(\beta)}{\sigma_1} \right) + \alpha \left( \frac{\log y_t - z_t(\gamma)}{\sigma_2} \right) = \varepsilon_t, \quad (14.44)$$

où les hypothèses sur  $u_{1t}$  et  $u_{2t}$  impliquent que  $\varepsilon_t$  est  $N(0,1)$ . Comme les modèles composites artificiels introduits dans la Section 11.3, ce dernier modèle ne peut pas être estimé parce que de nombreux paramètres seront en général non identifiés. Cependant, en suivant la procédure utilisée pour obtenir les tests en  $J$  et  $P$ , nous pouvons remplacer les paramètres du modèle qui n'est pas testé par des estimations. Ainsi, si nous désirons tester  $H_1$ , nous pouvons remplacer  $\gamma$  et  $\sigma_2$  par les estimations ML  $\hat{\gamma}$  et  $\hat{\sigma}_2$  de telle sorte que  $H_C$  devienne

$$H'_C: (1 - \alpha) \left( \frac{y_t - x_t(\beta)}{\sigma_1} \right) + \alpha \left( \frac{\log y_t - z_t(\hat{\gamma})}{\hat{\sigma}_2} \right) = \varepsilon_t.$$

Il est simple de tester  $H_1$  contre  $H'_C$  au moyen de la DLR:

$$\begin{bmatrix} \frac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{X}}_t & \frac{(y_t - \hat{x}_t)}{\hat{\sigma}_1} & \hat{z}_t - \log y_t \\ \mathbf{0} & -1 & \hat{\sigma}_1/y_t \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ s \\ a \end{bmatrix} + \text{résidus}, \quad (14.45)$$

où  $\hat{x}_t \equiv x_t(\hat{\beta})$ ,  $\hat{\mathbf{X}}_t \equiv \mathbf{X}_t(\hat{\beta})$ , et  $\hat{z}_t \equiv z_t(\hat{\gamma})$ . La DLR (14.45) est en fait une version simplifiée de la DLR que l'on obtient initialement. Tout d'abord,  $\hat{\sigma}_1$  fois le régresseur d'origine pour  $\sigma_1$  a été additionné au régresseur d'origine pour  $\alpha$ . Alors les régresseurs qui correspondent à  $\beta$  et à  $\sigma_1$  ont été multipliés par  $\hat{\sigma}_1$ , et les régresseurs qui correspondent à  $\alpha$  ont été multipliés par  $\hat{\sigma}_2$ . Aucune des ces modifications n'affecte le sous-espace engendré par les régresseurs, et par conséquent, aucun d'entre eux n'affecte les statistiques de test qu'on obtient. La dernière colonne de la matrice des régresseurs dans (14.45) est celle qui correspond à  $\alpha$ . Les autres colonnes devraient être orthogonales à la régressande par construction.

De façon similaire, si nous espérons tester  $H_2$ , nous pouvons remplacer  $\beta$  et  $\sigma_1$  par les estimations ML  $\hat{\beta}$  et  $\hat{\sigma}_1$  de telle sorte que  $H_C$  devienne

$$H''_C: (1 - \alpha) \left( \frac{y_t - x_t(\hat{\beta})}{\hat{\sigma}_1} \right) + \alpha \left( \frac{\log y_t - z_t(\gamma)}{\sigma_2} \right) = \varepsilon_t.$$

Il est alors simple de tester  $H_2$  contre  $H''_C$  au moyen de la DLR

$$\begin{bmatrix} \frac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{Z}}_t & \frac{\log y_t - \hat{z}_t}{\hat{\sigma}_2} & \hat{x}_t - y_t \\ \mathbf{0} & -1 & \hat{\sigma}_2 y_t \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ s \\ a \end{bmatrix} + \text{résidus}. \quad (14.46)$$

Encore une fois, ceci est une version simplifiée de la DLR que l'on obtient au début, et la dernière colonne de la matrice du régresseur est celle qui correspond à  $\alpha$ .

Les tests dont nous venons juste de discuter se généralisent bien évidemment très facilement aux modèles qui comprennent n'importe quelle sorte de

transformation de la variable dépendante, y compris les modèles de Box-Cox et d'autres modèles dans lesquels la transformation dépend d'un ou plusieurs paramètres inconnus. Pour plus de détails, consulter Davidson et MacKinnon (1984a). Il devrait être bien précisé que le modèle composite artificiel (14.44) est très arbitraire. Contrairement au modèle qui semble similaire pour les modèles de régression qui a été employé dans la Section 11.3, il ne fournit pas des tests asymptotiquement équivalents aux tests de Cox. De plus, peu de choses sont connues concernant les propriétés en échantillons finis des tests basés sur les DLR comme (14.45) et (14.46).

Une procédure finale qu'il est bon de mentionner est le **test  $P_E$**  suggéré par MacKinnon, White, et Davidson (1983). Il part aussi du modèle composite artificiel (14.44) mais ensuite il suit essentiellement l'approche du test d'Andrews de façon à obtenir une GNR qui teste seulement une direction de régression. Les régressions de test de Gauss-Newton pour le test  $P_E$  sont

$$y_t - \hat{x}_t = \hat{\mathbf{X}}_t \mathbf{b} + a(\hat{z}_t - \log \hat{x}_t) + \text{résidu} \quad (14.47)$$

pour le test de  $H_1$  et

$$\log y_t - \hat{z}_t = \hat{\mathbf{Z}}_t \mathbf{c} + d(\hat{x}_t - \exp \hat{z}_t) + \text{résidu} \quad (14.48)$$

pour le test de  $H_2$ . Les statistiques de test les plus simples à utiliser sont les  $t$  de Student pour  $a = 0$  dans (14.47) et  $d = 0$  dans (14.48). Comme le test d'Andrews, le test  $P_E$  manque très probablement de puissance, sauf quand la variance du DGP est très petite. Son tout premier avantage est que, contrairement aux tests basés sur la DLR, il sera asymptotiquement insensible aux caractéristiques de l'hypothèse de normalité.

## 14.7 LES AUTRES TRANSFORMATIONS

Les modèles basés sur la transformée de Box-Cox ne fonctionneront pas de façon adéquate à chaque fois. En particulier, le modèle de Box-Cox conventionnel n'est pas souvent très satisfaisant, pour des raisons que nous allons discuter. Dans cette section, nous discutons brièvement d'un nombre d'autres transformations qui peuvent être utiles dans certains cas. Nous n'en dirons pas beaucoup concernant les méthodes d'estimation et d'inférence pour ces modèles, sauf de noter qu'elles peuvent toutes être estimées par maximum de vraisemblance, en utilisant la DLR comme une partie de l'algorithme de maximisation, et que la DLR peut toujours être utilisée pour calculer des matrices de covariance et des statistiques de tests.

Un problème majeur avec le modèle de Box-Cox conventionnel est que le paramètre de transformation  $\lambda$  joue deux rôles différents: il modifie les propriétés des résidus, et change aussi la forme fonctionnelle de la fonction de régression. Par exemple, supposons que le DGP soit réellement un modèle de

régression linéaire à erreurs hétéroscédastiques qui ont une variance proportionnelle au carré de l'espérance de la variable dépendante :

$$y_t = \mathbf{X}_t \boldsymbol{\beta}_0 + u_t, \quad u_t \sim N(0, \sigma_0^2 (\mathbf{X}_t \boldsymbol{\beta}_0)^2), \quad (14.49)$$

où  $\sigma_0$  et  $\boldsymbol{\beta}_0$  désignent des valeurs sous le DGP. Si nous estimions un modèle de Box-Cox conventionnel en utilisant les données générées de cette manière, nous obtiendrions presque certainement une estimation de  $\lambda$  qui serait inférieure à l'unité, parce que ceci réduirait l'hétéroscédasticité dans les résidus. Ainsi, nous pourrions conclure de façon incorrecte qu'une spécification linéaire était inappropriée ou même qu'une spécification loglinéaire était inappropriée.

Le problème est que le paramètre de transformation dans le modèle de Box-Cox conventionnel affecte à la fois la forme de la fonction de régression et l'hétéroscédasticité des résidus. Une solution évidente est de permettre de façon explicite l'hétéroscédasticité, comme dans le modèle

$$B(y_t, \lambda) = \sum_{i=1}^k \beta_i B(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim N(0, \sigma^2 h(\mathbf{w}_t \boldsymbol{\delta})),$$

où  $h(\cdot)$  est une fonction scédastique,  $\mathbf{w}_t$  est un vecteur des observations des variables dépendantes, et  $\boldsymbol{\delta}$  est un vecteur des paramètres à estimer. Si on est tout d'abord intéressé par l'hétéroscédasticité de la forme qui apparaît dans (14.49), la fonction scédastique  $h(\mathbf{w}_t \boldsymbol{\delta})$  pourrait être spécifiée comme  $(\mathbf{X}_t \boldsymbol{\beta})^2$ . Consulter, parmi d'autres, Gaudry et Dagenais (1979), Lahiri et Egy (1981), et Tse (1984).

Une autre possibilité est de permettre qu'il y ait plus d'un paramètre de transformation, comme dans les modèles

$$B(y_t, \lambda) = \sum_{i=1}^k \beta_i B(X_{ti}, \phi) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t \quad \text{et} \quad (14.50)$$

$$B(y_t, \lambda) = B\left(\left(\sum_{i=1}^k \beta_i B(X_{ti}, \phi) + \sum_{j=1}^l \gamma_j Z_{tj}\right), \lambda\right) + u_t, \quad (14.51)$$

où, dans les deux cas,  $u_t$  est supposé  $N(0, \sigma^2)$ . Le premier de ces modèles est une généralisation évidente du modèle de Box-Cox conventionnel, et a été utilisé un certain nombre de fois en économétrie, parfois avec plus d'un paramètre  $\phi$ . Le second combine le modèle de Box-Cox conventionnel avec le modèle transformé des deux côtés, et n'a été utilisé dans aucun domaine. Dans les deux cas, le paramètre  $\phi$  affecte tout d'abord la forme fonctionnelle de la fonction de régression, tandis que le paramètre  $\lambda$  affecte tout d'abord les propriétés des aléas. Naturellement, il est loin d'être clair de savoir lequel des modèles (14.50), (14.51), et le modèle de Box-Cox conventionnel, sera le plus performant dans un cas donné.



Comme nous l'avons vu, la transformée de Box-Cox ne peut pas être appliquée aux variables qui peuvent prendre une valeur nulle ou négative. De nombreux auteurs, incluant John et Draper (1980) et Bickel et Doksum (1981), ont proposé des manières pour l'étendre de telle sorte qu'elle puisse être utilisée dans de tels cas. Par exemple, la proposition de Bickel-Doksum consiste à utiliser la transformation

$$\frac{\text{sign}(y)|y|^\lambda - 1}{\lambda} \quad (14.52)$$

à la place de la transformée de Box-Cox. Il est logiquement possible d'appliquer (14.52) aux variables qui peuvent prendre de petites valeurs (mais non nulles) et à celles qui prennent des valeurs négatives. Cependant, cette transformation ne comporte pas des propriétés particulièrement attrayantes; consulter Magee (1988). Quand  $\lambda$  est petit, elle a une pente extrêmement forte pour  $y$  proche de zéro. En plus, quand  $y < 0$ , (14.52) n'a pas de limite quand  $\lambda \rightarrow 0$ .

Il n'existe pas de raison de restreindre l'attention aux versions modifiées de la transformée de Box-Cox, puisque d'autres transformations peuvent bien être plus appropriées pour certains types de données. Par exemple, quand  $y_t$  est contrainte à rester comprise entre zéro et un, un modèle comme

$$y_t = \mathbf{X}_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim N(0, \sigma^2),$$

ne comporte réellement pas de sens, parce qu'il existe toujours une chance que  $u_t$  soit si grand que  $y_t$  tombe en dehors de l'intervalle 0-1. Dans un tel cas, il peut être souhaitable d'employer la transformation

$$\tau(y) = \log\left(\frac{y}{1-y}\right),$$

puisque  $\tau(y)$  peut varier entre moins l'infini et plus l'infini. Cette transformation ne comporte pas de paramètre inconnu, et ainsi ne nécessite pas que l'on quitte le cadre des modèles de régression; consulter Cox (1970).

Une famille intéressante de transformations a été étudiée par Burbidge, Magee, et Robb (1988) et MacKinnon et Magee (1990). Ces transformations ont la forme  $\theta(\alpha y)/\alpha$ , où la fonction  $\theta(\cdot)$  est croissante en ses arguments et possède les propriétés suivantes :

$$\theta(0) = 0; \quad \theta'(0) = 1; \quad \theta''(0) \neq 0. \quad (14.53)$$

Contrairement à la transformée de Box-Cox, cette transformation peut être appliquée aux variables d'un autre signe et aux variables nulles. Certaines fonctions  $\theta(\cdot)$  possèdent les propriétés (14.53). Une des plus simples est la fonction  $y + y^2$ , pour laquelle la transformation serait

$$\frac{\theta(\alpha y)}{\alpha} = y + \alpha y^2. \quad (14.54)$$

Evidemment, celle-ci sera une fonction convexe de  $y$  quand  $\alpha$  est une fonction positive et concave quand  $\alpha$  est négative. N'importe quelle transformation de la forme  $\theta(\alpha y)/\alpha$  qui satisfait (14.53) sera localement équivalente à (14.54), et ainsi nous voyons qu'un test de  $\alpha = 0$  peut être interprété comme un test contre n'importe quelle forme de non linéarité quadratique locale.

Pour cette famille de transformation, le modèle (14.04) deviendrait

$$\frac{\theta(\alpha y_t)}{\alpha} = x_t(\beta) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Il est facile de tester l'hypothèse nulle que  $\alpha = 0$  en utilisant une DLR très similaire à celle utilisée pour tester l'hypothèse nulle selon laquelle  $\lambda = 1$  dans un modèle de Box-Cox simple (14.04); consulter MacKinnon et Magee (1990) pour plus de détails. Ce test est sensible à plusieurs formes communes de mauvaises spécifications de modèle, qui incluent la non linéarité dans la fonction de régression, l'hétéroscédasticité, et l'asymétrie. Cela tend à être étroitement relié au test bien connu RESET; consulter la Section 6.5. On obtiendrait le test RESET si la transformation s'appliquait à  $x_t(\beta)$  à la place de  $y_t$ , comme dans le modèle

$$y_t = \frac{\theta(\alpha x_t(\beta))}{\alpha} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Puisqu'il s'agit simplement d'un modèle de régression non linéaire, un test pour  $\alpha = 0$  peut être basé sur une GNR. Il s'agit simplement du  $t$  de Student pour  $a = 0$  dans

$$y_t - x_t(\hat{\beta}) = \mathbf{X}_t(\hat{\beta})\mathbf{b} + ax_t^2(\hat{\beta}) + \text{résidu},$$

qui est une forme du test RESET.

## 14.8 CONCLUSION

À l'exception du modèle de Box-Cox conventionnel, les modèles qui comprennent des transformations de la variable dépendante ont été plutôt rarement utilisés en économétrie. Ceci est surprenant, parce qu'ils fournissent souvent une manière simple et peu coûteuse d'obtenir un modèle pour lequel les résidus se comportent bien et il existe une importante littérature les concernant en statistique, comprenant des ouvrages de McCullagh et Nelder (1983), Atkinson (1985), et Carroll et Ruppert (1988).

Nous avons vu dans ce chapitre qu'il n'est pas du tout difficile de traiter des modèles de ce type. Pourvu que l'on veuille supposer la normalité — et une telle hypothèse de distribution semble être nécessaire du moment que l'on quitte le cadre des modèles de régression — il est simple de les estimer par maximum de vraisemblance. La régression artificielle à longueur double

est extrêmement utile dans le contexte de ces modèles. Tout ce que l'on peut faire avec la régression de Gauss-Newton pour les modèles de régression non linéaire peut être réalisé avec la DLR pour les modèles qui comprennent des transformations de la variable dépendante. La régression OPG peut être utilisée à la place de la DLR, mais sera généralement moins performante.

## TERMES ET CONCEPTS

algorithmes de maximisation utilisant la DLR	régression artificielle (formulation générale)
facteurs Jacobiens	régressions linéaire contre loglinéaire
modèles de Box-Cox : conventionnel, simple, et transformé des deux côtés	termes Jacobiens
modèles autres que les régressions	tests non emboîtés
observations artificielles (pour DLR)	test $P_E$
régression artificielle à longueur double (DLR)	test RESET
	transformée de Box-Cox
	transformation non linéaire

# Chapitre 15

## Variables Dépendantes Limitées et Qualitatives

### 15.1 INTRODUCTION

Les modèles de régression supposent de manière implicite que la variable dépendante, peut-être après une transformation logarithmique ou autre, peut prendre n'importe quelle valeur sur la droite des réels. Bien que cette supposition ne soit pas strictement correcte pour les données économiques, elle est assez souvent raisonnable. Cependant, il s'agit d'une hypothèse acceptable lorsque la variable dépendante peut prendre n'importe quelle valeur spécifique de probabilité significativement supérieure à zéro. Les économistes ont fréquemment à faire à de tels cas. Les plus communément rencontrés sont les cas pour lesquels la variable dépendante peut prendre seulement deux valeurs. Par exemple, une personne peut faire partie de la population active ou non, un ménage peut être propriétaire ou locataire du logis où il vit, un débiteur peut faire défaut ou non à un prêt, un conducteur peut se déplacer pour son travail ou pour son loisir, et ainsi de suite. Ces cas constituent des exemples de **variables binaires dépendantes**.

Si nous désirons expliquer des variables économiques comme celles-ci dans un modèle économétrique, nous devons tenir compte de leur nature discrète. Les modèles de la sorte sont appelés **modèles à réponses qualitatives**, et sont habituellement estimés par la méthode du maximum de vraisemblance. Dans le cas le plus simple et le plus fréquent, la variable dépendante représente une ou deux alternatives. Elles sont codées de façon conventionnelle par 0 et 1, une convention qui se révèle être très pratique. Les modèles qui tentent d'expliquer les variables 0-1 sont souvent appelés **modèles à réponse binaire** ou, moins souvent, **modèles à choix binaire**. Ils sont très fréquemment employés en économie appliquée et dans de nombreux autres domaines où s'applique l'économétrie, comme les exemples précédents servent à l'illustrer.

Les modèles de régression sont aussi inappropriés pour traiter les modèles comprenant des **variables dépendantes limitées**, pour lesquels il existe une grande quantité de variétés. Parfois une variable dépendante peut être continue sur un ou plusieurs intervalles de la droite des réels mais peut prendre une ou plusieurs valeurs avec une probabilité finie. Par exemple, les

dépenses de consommation portant sur certaines catégories de biens et services sont généralement contraintes à être non négatives. Ainsi, si nous observons les dépenses portant sur une certaine catégorie pour un échantillon de biens ménagers, il est très probable que ces dépenses seront nulles pour certains biens ménagers et positives pour d'autres. Comme il existe une probabilité positive qu'une valeur particulière, zéro, se présente dans les données les modèles de régression ne sont pas appropriés pour ce type de données. Un autre type de modèle à variables dépendantes limitées survient quand seulement certains résultats (tels que les résultats positifs dans cet exemple) sont observés. Ceci signifie que l'échantillon ne sera pas aléatoire.

Dans ce chapitre, nous traitons à la fois les modèles à réponse qualitative et les modèles à variables dépendantes limitées. Il s'agit d'un domaine dans lequel il y a eu une énorme quantité de recherche durant les 20 dernières années, et c'est pourquoi notre traitement couvre seulement quelques uns des modèles les plus basiques. Nous nous concentrerons tout d'abord sur les modèles à réponse binaire, parce qu'ils sont à la fois les modèles les plus simples et les plus fréquents. Ils seront discutés dans les trois prochaines sections. Ensuite, dans la Section 15.5, nous discuterons brièvement des modèles à réponses qualitatives pour les cas comprenant plus de deux réponses différentes. Finalement, dans les trois dernières sections, nous portons notre attention sur certains des modèles les plus simples qui concernent les variables dépendantes limitées.

## 15.2 LES MODÈLES À RÉPONSE BINAIRE

Dans un modèle à réponse binaire, la valeur de la variable dépendante  $y_t$  peut prendre seulement deux valeurs, 1 et 0, qui indiquent si un certain événement se produit ou pas. Nous pouvons proposer que  $y_t = 1$  indique que l'événement s'est produit pour l'observation  $t$  et que  $y_t = 0$  indique que l'événement ne s'est pas produit. Soit  $P_t$  la probabilité (conditionnelle) que l'événement se soit produit. Ainsi un modèle à réponse binaire essaie vraiment de modéliser la probabilité  $P_t$  conditionnelle à un certain ensemble d'informations, disons  $\Omega_t$ , qui se compose de variables prédéterminées et exogènes. Ainsi la spécification de  $y_t$  qui est soit 0 soit 1 est très commode, parce que la probabilité  $P_t$  constitue alors simplement l'espérance de  $y_t$  conditionnelle à l'ensemble d'information  $\Omega_t$ :

$$P_t \equiv \Pr(y_t = 1 | \Omega_t) = E(y_t | \Omega_t).$$

L'objectif d'un modèle à réponse binaire est de modéliser cette espérance conditionnelle.

Partant de cette perspective, il est clair qu'un modèle de régression linéaire est moins bien adapté qu'un modèle à réponse binaire. Supposons que  $\mathbf{X}_t$  désigne un vecteur ligne de dimension  $k$  des variables qui appartiennent à l'ensemble d'information  $\Omega_t$ , qui inclut un terme constant ou l'équivalent.

Alors un modèle de régression linéaire spécifierait  $E(y_t | \Omega_t)$  pour  $\mathbf{X}_t\boldsymbol{\beta}$ . Mais  $E(y_t | \Omega_t)$  est une probabilité, et les probabilités doivent être comprises entre 0 et 1. La quantité  $\mathbf{X}_t\boldsymbol{\beta}$  n'est pas contrainte de la sorte et par conséquent, elle ne peut pas être interprétée comme une probabilité. Néanmoins, beaucoup de travaux empiriques (pour la plupart plus anciens) utilisent simplement les OLS pour estimer ce qui est appelé (plutôt de manière maladroite) le **modèle de probabilité linéaire**,<sup>1</sup> qui est le modèle

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t.$$

Etant donné que des modèles bien meilleurs sont disponibles et qu'il est facile de les estimer en utilisant la technologie informatique moderne, ce type de modèle n'est presque pas recommandable. Même s'il arrive que  $\mathbf{X}_t\boldsymbol{\beta}$  soit compris entre 0 et 1 pour un  $\boldsymbol{\beta}$  quelconque et toutes les observations dans un échantillon particulier, il est impossible de contraindre  $\mathbf{X}_t\boldsymbol{\beta}$  à rester dans cet intervalle pour toutes les valeurs possibles de  $\mathbf{X}_t$ , à moins que les valeurs prises par les variables indépendantes soient limitées d'une certaine manière (par exemple, elles peuvent toutes être des variables muettes). Ainsi le modèle de probabilité linéaire ne constitue pas un moyen judicieux pour modéliser les probabilités conditionnelles.

Plusieurs modèles à réponse binaire pertinents sont disponibles et sont très faciles à traiter. La subtilité consiste à utiliser une **fonction de transformation**  $F(x)$  qui comporte les propriétés

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad \text{et} \quad (15.01)$$

$$f(x) \equiv \frac{\partial F(x)}{\partial x} > 0. \quad (15.02)$$

Ainsi  $F(x)$  est une fonction monotone croissante qui s'applique de la droite des réels vers l'intervalle 0-1. Certaines fonctions de distribution cumulées comportent ces propriétés, et nous discuterons brièvement de certains exemples spécifiques. En utilisant des spécifications variées pour la fonction de transformation, nous pouvons modéliser l'espérance conditionnelle de  $y_t$  de plusieurs manières.

Les modèles à réponse binaire dont nous discuterons se composent d'une fonction de transformation  $F(x)$  appliquée à une **fonction indice** qui dépend des variables indépendantes et des paramètres du modèle. Une fonction indice est simplement une fonction qui comporte les propriétés d'une fonction de régression, soit linéaire soit non linéaire. Ainsi une spécification très générale d'un modèle à réponse binaire est

$$E(y_t | \Omega_t) = F(h(\mathbf{X}_t, \boldsymbol{\beta})),$$

<sup>1</sup> Consulter, par exemple, Bowen et Finegan (1969).

où  $h(\mathbf{X}_t, \boldsymbol{\beta})$  est la fonction indice. Une spécification plus restrictive, mais plus fréquente, est

$$E(y_t | \Omega_t) = F(\mathbf{X}_t \boldsymbol{\beta}). \quad (15.03)$$

Dans ce cas, la fonction indice  $\mathbf{X}_t \boldsymbol{\beta}$  est linéaire et  $E(y_t | \Omega_t)$  est simplement une transformation non linéaire. Bien que  $\mathbf{X}_t \boldsymbol{\beta}$  puisse en principe prendre n'importe quelle valeur sur la droite des réels,  $F(\mathbf{X}_t \boldsymbol{\beta})$  doit être comprise entre 0 et 1 d'après la propriété (15.01).

Parce que  $F(\cdot)$  est une fonction non linéaire, les changements dans les valeurs de  $X_{ti}$ , qui sont les éléments de  $\mathbf{X}_t$ , affectent nécessairement  $E(y_t | \Omega_t)$  d'une manière non linéaire. De façon plus spécifique, quand  $P_t \equiv E(y_t | \Omega_t)$  est fournie par (15.03), sa dérivée par rapport à  $X_{ti}$  est

$$\frac{\partial P_t}{\partial X_{ti}} = \frac{\partial F(\mathbf{X}_t \boldsymbol{\beta})}{\partial X_{ti}} = f(\mathbf{X}_t \boldsymbol{\beta}) \beta_i. \quad (15.04)$$

Pour les fonctions de transformation qui sont presque toujours employées,  $f(\mathbf{X}_t \boldsymbol{\beta})$  atteint son maximum en zéro et décroît ensuite quand  $\mathbf{X}_t \boldsymbol{\beta}$  s'éloigne de zéro. Ainsi, (15.04) nous indique que l'effet sur  $P_t$  d'un changement d'une des variables dépendantes est maximum lorsque  $P_t = .5$  et minimum lorsque  $P_t$  est proche de 0 ou 1.

Quand les modèles à réponse binaire sont utilisés dans un travail appliqué, la fonction indice linéaire  $\mathbf{X}_t \boldsymbol{\beta}$  est presque toujours employée, parmi une des deux spécifications pour  $F(\cdot)$ . Les modèles qui en résultent sont appelés **modèle probit** et **modèle logit**. Pour le modèle probit, la fonction de transformation  $F(x)$  est la fonction de distribution cumulée de la loi normale standard

$$\Phi(x) \equiv \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}X^2\right) dX.$$

Comme  $\Phi(x)$  est une fonction de répartition, elle satisfait automatiquement les conditions (15.01) et (15.02). Le modèle probit peut être écrit comme

$$P_t \equiv E(y_t | \Omega_t) = \Phi(\mathbf{X}_t \boldsymbol{\beta}).$$

Bien qu'il n'existe aucune expression bornée pour  $\Phi(x)$ , elle est facilement évaluée numériquement, et sa dérivée première est naturellement la fonction de densité de la loi normale standard

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

Le modèle probit peut provenir d'un modèle comprenant une variable  $y_t^*$  non observée, ou **latente**. Supposons que

$$y_t^* = \mathbf{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, 1). \quad (15.05)$$

Nous observons seulement le signe de  $y_t^*$ , qui détermine la valeur de la variable binaire observée  $y_t$  selon la relation

$$y_t = 1 \text{ si } y_t^* > 0 \quad \text{et} \quad y_t = 0 \text{ si } y_t^* \leq 0. \quad (15.06)$$

Par exemple, nous pourrions imaginer que  $y_t^*$  est un indice de l'utilité (nette) obtenue de certaine action. Si l'action fournit une utilité positive, elle sera retenue; et ne le sera pas si l'action fournit une utilité négative ou nulle. Comme nous observons seulement si l'action est ou n'est pas retenue, nous observons seulement le signe de  $y_t^*$ . De ce fait, nous pouvons normaliser la variance de  $u_t$  à l'unité. Si  $u_t$  avait réellement une autre variance quelconque, disons  $\sigma^2$ , la division de  $y_t^*$ ,  $\beta$ , et  $u_t$  par  $\sigma$  fournirait un modèle d'observation identique à celui d'origine.

Maintenant, nous pouvons nous demander à quoi correspond la probabilité  $y_t = 1$ . Certaines manipulations simples fournissent

$$\begin{aligned} \Pr(y_t = 1) &= \Pr(y_t^* > 0) = \Pr(\mathbf{X}_t\beta + u_t > 0) \\ &= 1 - \Pr(u_t \leq -\mathbf{X}_t\beta) = 1 - \Phi(-\mathbf{X}_t\beta) = \Phi(\mathbf{X}_t\beta). \end{aligned} \quad (15.07)$$

La dernière égalité dans (15.07) utilise le fait que la fonction de densité normale est symétrique par rapport à zéro. Le résultat final,  $\Phi(\mathbf{X}_t\beta)$ , est simplement la probabilité que nous obtiendrions en remplaçant  $F(\cdot)$  par  $\Phi(\cdot)$  dans (15.03). Ainsi nous avons dérivé le modèle probit à partir du **modèle à variable latente** composé de (15.05) et (15.06). Le fait que le modèle probit puisse être dérivé de cette manière constitue une de ses caractéristiques les plus attrayantes.

Le modèle logit est très similaire au modèle probit mais possède un nombre de caractéristiques qui le rendent plus facile à utiliser. Pour le modèle logit, la fonction  $F(x)$  est la **fonction logistique**

$$\Lambda(x) \equiv (1 + e^{-x})^{-1} = \frac{e^x}{1 + e^x},$$

qui a comme dérivée première

$$\lambda(x) \equiv \frac{e^x}{(1 + e^x)^2} = \Lambda(x)\Lambda(-x).$$

La seconde égalité se révélera très utile plus tard. Le modèle est plus facilement dérivé en supposant que

$$\log\left(\frac{P_t}{1 - P_t}\right) = \mathbf{X}_t\beta,$$

qui indique que le logarithme des probabilités est égal à  $\mathbf{X}_t\beta$ . En résolvant par rapport à  $P_t$ , nous trouvons que

$$P_t = \frac{\exp(\mathbf{X}_t\beta)}{1 + \exp(\mathbf{X}_t\beta)} = (1 + \exp(-\mathbf{X}_t\beta))^{-1} = \Lambda(\mathbf{X}_t\beta).$$



Il est aussi possible de dériver le modèle logit à partir d'un modèle à variable latente comme (15.05) et (15.06) mais avec des erreurs qui suivent une distribution à valeur extrême au lieu d'une normale; consulter, parmi d'autres, Domencich et McFadden (1975), McFadden (1984), et Train (1986).

Dans la pratique, les modèles logit et probit tendent à fournir des résultats assez similaires. Dans la plupart des cas, la seule différence réelle entre eux réside dans la manière dont les éléments de  $\beta$  sont gradués. Cette différence dans la graduation survient parce que la variance de la distribution lorsque la fonction logistique est la fonction de répartition est  $\pi^2/3$ , tandis que celle de la loi normale standard est naturellement égale à l'unité. Ainsi les estimations logit tendent toutes à être supérieures aux estimations probit, habituellement d'un facteur juste inférieur à  $\pi/\sqrt{3}$ .<sup>2</sup> La Figure 15.1 illustre les fonctions de répartition des loi normale standard, logistique, et logistique regraduée pour obtenir une variance unitaire. La similitude entre la fonction de répartition de la loi normale et la fonction logistique regraduée est frappante.

Au vu de leurs propriétés similaires, il est peut-être curieux qu'à la fois les modèles logit et probit continuent à être largement employés, tandis que des modèles véritablement différents des deux précédents sont très rarement rencontrés. Il existe autant de manières de spécifier de tels modèles qu'il existe de choix plausibles pour la fonction de transformation  $F(x)$ . Par exemple, un tel choix est

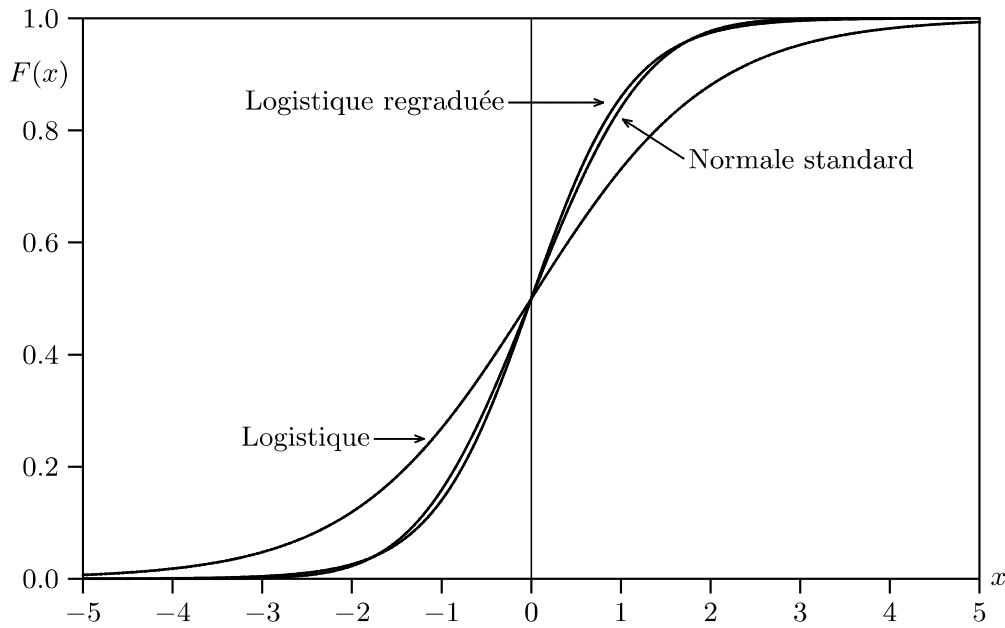
$$F(x) = \pi^{-1} \arctan(x) + \frac{1}{2}. \quad (15.08)$$

Comme il s'agit de la fonction de répartition de Cauchy, sa dérivée est

$$f(x) = \frac{1}{\pi(1+x^2)},$$

qui est la densité de Cauchy (consulter la Section 4.6). Comme le comportement de la fonction de distribution de Cauchy dans les queues est très différent de celui d'autres fonctions de distribution comme  $\Phi(x)$  ou  $\Lambda(x)$ , il existe au moins la possibilité qu'un modèle à réponse binaire basé sur (15.08) soit plus ou moins performant qu'un modèle logit ou probit. D'un autre côté, il existe une infime probabilité pour que ces deux modèles fournissent des résultats qui diffèrent de manière significative, à moins que la taille de l'échantillon soit en fait très importante.

<sup>2</sup> Amemiya (1981) suggère que 1.6, plutôt que  $\pi/\sqrt{3} \cong 1.81$  peut être une meilleure estimation du facteur par lequel les estimations logit tendent à excéder les estimations probit. Greene (1990a) remarque aussi qu'une justification pour cette régularité est que  $\phi(0)/\lambda(0) \cong 1.6$ . Souvenons-nous de (15.04) que les dérivées de  $P_t$  par rapport à  $X_{ti}$  sont égales à  $f(\mathbf{X}_t\beta)\beta_i$ . Si  $\mathbf{X}_t\beta$  est approximativement nul en moyenne et que les modèles logit et probit prédisent le même effet sur  $P_t$  pour une variation donnée des  $X_{ti}$ , alors les coefficients pour le modèle logit doivent être approximativement 1.6 fois ceux du modèle probit. On peut s'attendre à ce que cette approximation s'adapte moins bien quand la valeur moyenne de  $P_t$  est loin de .5.



**Figure 15.1** Trois choix possibles de  $F(x)$

Les trois choix pour  $F(\cdot)$  que nous avons discutés sont symétriques par rapport à zéro. Cela signifie qu'elles ont la propriété que  $1 - F(x) = F(-x)$ , qui implique que  $f(x) = f(-x)$ . Il s'agit parfois d'une propriété commode, mais il n'existe pas de raison a priori pour s'y tenir. Les choix pour  $F(\cdot)$  qui ne possèdent pas cette propriété fourniront potentiellement des résultats très différents de ceux produits par les modèles logit et probit. Une manière d'obtenir le même effet consiste à spécifier le modèle comme

$$E(y_t | \Omega_t) = F(h(\mathbf{X}_t\boldsymbol{\beta})),$$

où  $F(\cdot)$  est  $\Phi(\cdot)$  ou  $\Lambda(\cdot)$ , et  $h(\cdot)$  est une transformation non linéaire. Ceci suggère une façon de tester la validité de l'hypothèse de symétrie oblique, sujet que nous aborderons dans la Section 15.4.

### 15.3 ESTIMATION DES MODÈLES À RÉPONSE BINAIRE

A présent, le moyen de loin le plus communément employé pour estimer les modèles à réponse binaire est l'utilisation de la méthode du maximum de vraisemblance. Nous limiterons notre attention à cette méthode et supposerons, pour simplifier, que la fonction indice est simplement  $\mathbf{X}_t\boldsymbol{\beta}$ . Ensuite, selon le modèle à réponse binaire (15.03),  $F(\mathbf{X}_t\boldsymbol{\beta})$  est la probabilité que  $y_t = 1$  et  $1 - F(\mathbf{X}_t\boldsymbol{\beta})$  est la probabilité que  $y_t = 0$ . Ainsi, si  $y_t = 1$ , la contribution au logarithme de la fonction de vraisemblance pour l'observation  $t$  est

$\log(F(\mathbf{X}_t\boldsymbol{\beta}))$ , tandis que si  $y_t = 0$ , la contribution est  $\log(1 - F(\mathbf{X}_t\boldsymbol{\beta}))$ . En conséquence, la fonction de vraisemblance est

$$\ell(\mathbf{y}, \boldsymbol{\beta}) = \sum_{t=1}^n \left( y_t \log(F(\mathbf{X}_t\boldsymbol{\beta})) + (1 - y_t) \log(1 - F(\mathbf{X}_t\boldsymbol{\beta})) \right). \quad (15.09)$$

Cette fonction est globalement concave à chaque fois que  $\log(F(x))$  et  $\log(1 - F(x))$  sont des fonctions concaves de l'argument  $x$ ; consulter Pratt (1981). Cette condition est satisfaite par de nombreux modèles à réponse binaire, incluant les modèles logit et probit. Par conséquent, les fonctions de logvraisemblance pour ces modèles sont très faciles à maximiser numériquement.<sup>3</sup>

Les conditions du premier ordre pour un maximum de (15.09) sont

$$\sum_{t=1}^n \frac{(y_t - \hat{F}_t) \hat{f}_t X_{ti}}{\hat{F}_t(1 - \hat{F}_t)} = 0, \quad i = 1, \dots, k, \quad (15.10)$$

où  $\hat{F}_t \equiv F(\mathbf{X}_t\hat{\boldsymbol{\beta}})$  et  $\hat{f}_t \equiv f(\mathbf{X}_t\hat{\boldsymbol{\beta}})$ , avec  $\hat{\boldsymbol{\beta}}$  qui désigne le vecteur des estimations ML. Toutes les fois que la fonction de logvraisemblance est globalement concave, ces conditions du premier ordre définissent un maximum unique si elles sont tout à fait satisfaites. Nous pouvons vérifier que les modèles logit, probit, et de nombreux autres modèles à réponse binaire satisfont les conditions de régularité nécessaires pour que les estimations  $\hat{\boldsymbol{\beta}}$  soient convergentes et asymptotiquement normales, avec une matrice de covariance asymptotique donnée par l'inverse de la matrice d'information selon la façon habituelle. Consulter, par exemple, Gouriéroux et Monfort (1981). Dans le cas du modèle logit, les conditions du premier ordre (15.10) se simplifient

$$\sum_{t=1}^n (y_t - \Lambda(\mathbf{X}_t\hat{\boldsymbol{\beta}})) X_{ti} = 0, \quad i = 1, \dots, k,$$

parce que  $\lambda(x) = \Lambda(x)(1 - \Lambda(x))$ . Notons que les conditions (15.10) ressemblent aux conditions du premier ordre de l'estimation par moindres carrés pondérés du modèle de régression non linéaire

$$y_t = F(\mathbf{X}_t\boldsymbol{\beta}) + e_t, \quad (15.11)$$

avec des poids donnés par

$$\left( F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta})) \right)^{-1/2}.$$

<sup>3</sup> Dans le cas usuel, où  $F(\cdot)$  est symétrique-oblique, il est plus judicieux d'évaluer  $\log(F(-\mathbf{X}_t\boldsymbol{\beta}))$  plutôt que  $\log(1 - F(\mathbf{X}_t\boldsymbol{\beta}))$  lors de l'écriture de programmes informatiques. Ceci évite le risque que  $1 - F(\mathbf{X}_t\boldsymbol{\beta})$  soit évalué de manière très imprécise lorsque  $F(\mathbf{X}_t\boldsymbol{\beta})$  est très proche de l'unité. Bien que  $F(\cdot)$  ne nécessite pas d'être symétrique-oblique, nous retiendrons la notation la plus générale.

Cela est logique du fait que la variance de l'aléa dans (15.11) est

$$\begin{aligned} E(e_t^2) &= E(y_t - F(\mathbf{X}_t\boldsymbol{\beta}))^2 \\ &= F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta}))^2 + (1 - F(\mathbf{X}_t\boldsymbol{\beta}))(F(\mathbf{X}_t\boldsymbol{\beta}))^2 \\ &= F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta})). \end{aligned}$$

Ainsi, une manière d'obtenir des estimations ML de n'importe quel modèle à réponse binaire consiste à appliquer par itérations les moindres carrés non linéaires repondérés à (15.11) ou à tout modèle de régression non linéaire approprié si la fonction indice n'est pas  $\mathbf{X}_t\boldsymbol{\beta}$ . Cependant, pour la plupart des modèles, cette stratégie ne constitue pas la meilleure approche. Une approche adéquate est exposée dans la prochaine section.

Comme le ML est équivalent à une forme de NLS pondérés pour les modèles à réponse binaire, il est évident que la matrice de covariance asymptotique pour  $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$  doit être

$$\left(\frac{1}{n}\mathbf{X}^\top\boldsymbol{\Psi}(\boldsymbol{\beta}_0)\mathbf{X}\right)^{-1},$$

où  $\mathbf{X}$  est une matrice de dimension  $n \times k$  avec comme ligne type  $\mathbf{X}_t$  et comme élément type  $X_{ti}$ , et  $\boldsymbol{\Psi}(\boldsymbol{\beta})$  est une matrice diagonale avec comme élément diagonal type

$$\Psi(\mathbf{X}_t\boldsymbol{\beta}) = \frac{f^2(\mathbf{X}_t\boldsymbol{\beta})}{F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta}))}. \quad (15.12)$$

Le numérateur reflète le fait que la dérivée de  $F(\mathbf{X}_t\boldsymbol{\beta})$  par rapport à  $\beta_i$  est  $f(\mathbf{X}_t\boldsymbol{\beta})X_{ti}$ , et le dénominateur est simplement la variance de  $e_t$  dans (15.11). Dans le cas du modèle logit,  $\Psi(\mathbf{X}_t\boldsymbol{\beta})$  se simplifie en  $\lambda(\mathbf{X}_t\boldsymbol{\beta})$ .

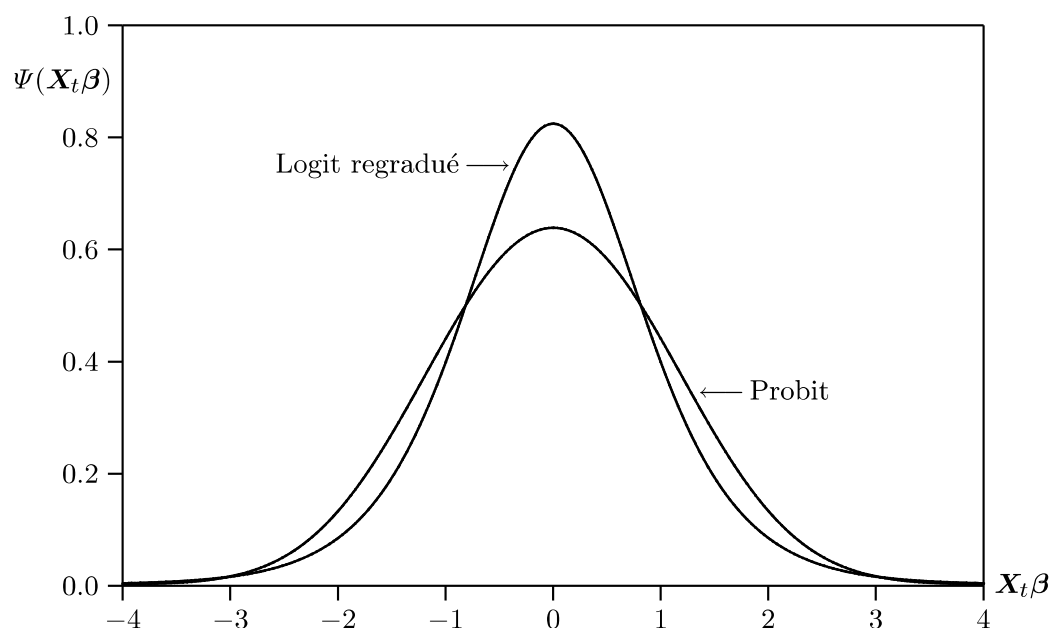
Cette matrice de covariance asymptotique peut aussi être obtenue en prenant l'inverse de la matrice d'information. Comme d'habitude, celle-ci est égale à l'espérance de l'opposé de  $n^{-1}$  fois la matrice Hessienne mais également à l'espérance du produit extérieur du gradient. La matrice d'information est simplement

$$\mathcal{I}(\boldsymbol{\beta}) \equiv \frac{1}{n}\mathbf{X}^\top\boldsymbol{\Psi}(\boldsymbol{\beta})\mathbf{X}, \quad (15.13)$$

où  $\boldsymbol{\Psi}(\boldsymbol{\beta})$  est définie par (15.12). Par exemple, à partir de (15.10) il est aisé de voir que l'élément type de la matrice  $n^{-1}\mathbf{G}^\top(\boldsymbol{\beta})\mathbf{G}(\boldsymbol{\beta})$ , où  $\mathbf{G}(\boldsymbol{\beta})$  est la matrice CG, est

$$\frac{1}{n} \sum_{t=1}^n \left( \frac{(y_t - F(\mathbf{X}_t\boldsymbol{\beta}))f(\mathbf{X}_t\boldsymbol{\beta})}{F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta}))} \right)^2 X_{ti}X_{tj}.$$

Montrer que l'espérance de cette expression est un élément type de la matrice d'information (15.13) constitue un bon exercice.



**Figure 15.2** Les poids pour les modèles probit et logit regradué

Décéler l'analogie entre les estimations provenant d'un modèle à réponse binaire et les estimations par moindres carrés pondérés est très révélateur. Dans le cas des moindres carrés, chaque observation est pondérée par un poids égal quand la matrice d'information est formée. Dans le cas à réponse binaire, d'un autre côté, certaines observations sont pondérées beaucoup plus que d'autres, parce que les poids  $\Psi(\mathbf{X}_t\boldsymbol{\beta})$  définis dans (15.12) peuvent différer fortement. Si on graphé ces pondérations comme une fonction de  $\mathbf{X}_t\boldsymbol{\beta}$  pour les modèles probit ou logit, nous trouvons que le poids maximum sera associé aux observations pour lesquelles  $\mathbf{X}_t\boldsymbol{\beta} = 0$ , ce qui implique que  $P_t = .5$ , tandis qu'un poids relativement faible sera associé aux observations pour lesquelles  $P_t$  est proche de 0 ou 1. Ceci est logique car lorsque  $P_t$  est proche de 0 ou 1, un changement dans  $\boldsymbol{\beta}$  aura un faible impact sur  $P_t$ , tandis que lorsque  $P_t$  est proche de .5, un changement aura un effet beaucoup plus important. Par conséquent les observations du dernier type fournissent beaucoup plus d'information que les observations du premier type.

Dans la Figure 15.2, les pondérations (15.12) sont graphées pour les cas probit et logit, (la dernière a été regraduée pour avoir une variance unitaire) comme des fonctions de l'indice  $\mathbf{X}_t\boldsymbol{\beta}$ . Notons que les différences entre ces deux modèles sont plus frappantes qu'elles ne le furent dans la Figure 15.1. Le modèle logit associe plus de poids aux observations pour lesquelles  $\mathbf{X}_t\boldsymbol{\beta}$  est proche ou loin de zéro, tandis que le modèle probit associe des poids plus importants aux observations pour lesquelles  $\mathbf{X}_t\boldsymbol{\beta}$  prend des valeurs intermédiaires (approximativement, entre 0.8 et 3.0). Cependant, les

différences qui sont apparentes dans la figure semblent rarement prendre plus d'importance dans la pratique.

Comme nous l'avons vu, nous pouvons penser qu'une variable dépendante binaire provienne d'un modèle à variable latente tel que celui donné par (15.05) et (15.06). Il est intéressant de se demander quel est le degré d'efficacité perdu par la variable latente non observable. Manifestement, quelque chose doit être perdu, parce qu'une variable binaire telle que  $y_t$  doit fournir moins d'information qu'une variable continue telle que  $y_t^*$ . La matrice de covariance pour les estimations OLS de  $\beta$  dans (15.05) est  $(\mathbf{X}^\top \mathbf{X})^{-1}$ ; rappelons que la variance d'erreur est normalisée à l'unité. Par contraste, la matrice de covariance pour les estimations probit de  $\beta$  est  $(\mathbf{X}^\top \Psi(\beta) \mathbf{X})^{-1}$ , où  $\Psi(\beta)$  était définie par (15.12). La valeur maximale pour  $\Psi(\mathbf{X}_t \beta)$  est atteinte quand  $P_t = .5$ . Dans le cas probit, cette valeur est 0.6366. Par conséquent, dans le meilleur cas possible, lorsque les données sont telles que  $P_t = .5$  pour tout  $t$ , la matrice de covariance pour les estimations probit sera égale à 1.57 ( $\cong 1/0.6366$ ) fois la matrice de covariance des OLS. Dans la pratique, naturellement, cette borne supérieure n'est probablement pas atteinte, et les estimations probit peuvent être beaucoup moins efficaces que ne le seraient les estimations OLS, qui utilisent la variable latente, en particulier lorsque  $P_t$  est proche de 0 ou 1 pour une partie importante de l'échantillon.

Un problème pratique avec les modèles à réponse binaire est que les conditions du premier ordre (15.10) n'ont pas nécessairement de solution finie. Ceci peut survenir quand l'ensemble des données ne fournit pas suffisamment d'information pour identifier tous les paramètres. Supposons qu'il existe une quelconque combinaison linéaire des variables indépendantes, disons  $z_t \equiv \mathbf{X}_t \beta^*$ , telle que

$$y_t = 0 \text{ pour } z_t \leq 0, \text{ et}$$

$$y_t = 1 \text{ pour } z_t > 0.$$

Alors il sera possible de faire tendre  $\ell(\mathbf{y}, \beta)$  vers zéro en posant  $\beta = \alpha \beta^*$  et en laissant  $\alpha \rightarrow \infty$ . Ceci garantira que  $F(\mathbf{X}_t \beta) \rightarrow 0$  pour toutes les observations où  $y_t = 0$  et  $F(\mathbf{X}_t \beta) \rightarrow 1$  pour toutes les observations où  $y_t = 1$ . La valeur de la fonction de logvraisemblance (15.09) tendra donc vers zéro quand  $\alpha \rightarrow \infty$ . Mais zéro est évidemment une borne supérieure pour cette valeur. Donc, dans de telles circonstances, les paramètres  $\beta$  ne sont pas identifiés sur l'espace paramétrique non compact  $\mathbb{R}^k$  au sens de la Définition 8.1, et nous ne pouvons pas obtenir des estimations pertinentes de  $\beta$ ; consulter Albert et Anderson (1984).

Quand  $z_t$  est simplement une combinaison linéaire du terme constant et d'une seule variable indépendante, cette dernière est souvent appelée **classificatrice parfaite**, parce que les  $y_t$  peuvent être classées en 0 ou 1, une fois la valeur de la variable connue. Par exemple, considérons le DGP

$$\begin{aligned} y_t^* &= x_t + u_t, & u_t &\sim \text{NID}(0, 1); \\ y_t &= 1 \text{ si } y_t^* > 0 \quad \text{et} \quad y_t = 0 \text{ si } y_t^* \leq 0. \end{aligned} \tag{15.14}$$

Pour ce DGP, il semblerait judicieux d'estimer le modèle probit

$$E(y_t | x_t) = \Phi(\beta_0 + \beta_1 x_t). \quad (15.15)$$

Mais supposons que, dans l'exemple,  $x_t$  soit toujours un nombre inférieur à  $-4$  ou supérieur à  $+4$ . Quand  $x_t$  est inférieur à  $-4$ , il est presque certain (la probabilité est supérieure à 0.99997) que  $y_t$  sera 0, et quand  $x_t$  est supérieure à  $+4$ , il est presque certain que  $y_t$  sera 1. Ainsi, à moins que la taille de l'échantillon soit très grande, il est peu probable qu'il y ait des observations pour lesquelles  $x_t < 0$  et  $y_t = 1$  ou des observations pour lesquelles  $x_t > 0$  et  $y_t = 0$ . En l'absence de telles observations, la variable  $x_t$  sera une classificatrice parfaite, et il sera impossible d'obtenir des estimations correctes des paramètres de (15.14). Quel que soit l'algorithme de maximisation utilisé, il essaiera simplement de rendre  $\hat{\beta}_1$  aussi grand que possible.

Bien que cet exemple soit extrême, des problèmes similaires sont susceptibles de survenir lorsque l'ajustement du modèle est très bon et la taille de l'échantillon est petite. Il existera une classificatrice parfaite quand il y a un hyperplan séparateur dans l'espace des explicatives tel que toutes les observations pour lesquelles  $y_t = 0$  se situent de l'un côté et toutes celles pour lesquelles  $y_t = 1$  de l'autre. Ce cas de figure est probable si l'ajustement est bon et il n'y a que peu d'observations avec  $y_t = 0$ , ou peu avec  $y_t = 1$ . Il se peut néanmoins que des estimations ML puissent se calculer même quand  $n$  n'est pas plus grand que  $k + 1$  et il n'y a qu'une seule observation avec soit  $y_t = 0$  soit  $y_t = 1$ .

Dans les modèles de régression, il est commun de tester l'hypothèse que toutes les pentes sont nulles en utilisant un test en  $F$ . Pour les modèles à réponse binaire, la même hypothèse peut facilement être testée en utilisant un test du ratio de vraisemblance. Un modèle avec un terme constant peut être écrit comme

$$E(y_t | \Omega_t) = F(\beta_1 + \mathbf{X}_{2t}\beta_2), \quad (15.16)$$

où  $\mathbf{X}_{2t}$  se compose de  $\mathbf{X}_t$  sans le terme constant et  $\beta_2$  est un vecteur de dimension  $(k - 1)$ . Sous l'hypothèse nulle que  $\beta_2 = \mathbf{0}$ , (15.16) devient

$$E(y_t | \Omega_t) = F(\beta_1) = E(y_t).$$

Ceci indique simplement que l'espérance conditionnelle de  $y_t$  est égale à son espérance non conditionnelle, qui peut être estimée par  $\bar{y}$ . Par conséquent, si  $\bar{\beta}_1$  désigne l'estimation de  $\beta_1$ ,  $\bar{y} = F(\bar{\beta}_1)$ . A partir de (15.09), il est aisé de voir que la valeur de la fonction de logvraisemblance sous l'hypothèse nulle est

$$\ell(\mathbf{y}, \bar{\beta}_1, \mathbf{0}) = n \bar{y} \log(\bar{y}) + n(1 - \bar{y}) \log(1 - \bar{y}). \quad (15.17)$$

Le double de la différence entre la valeur non contrainte  $\ell(\mathbf{y}, \hat{\beta}_1, \hat{\beta}_2)$  et la valeur contrainte  $\ell(\mathbf{y}, \bar{\beta}_1, \mathbf{0})$  constitue une statistique de test LR qui sera asymptotiquement distribuée suivant une  $\chi^2(k - 1)$ . Comme le membre de droite de

(15.17) est très facile à calculer, la statistique de test l'est également. Cependant, nous discuterons dans la prochaine section d'une statistique de test encore plus facile à calculer.

De nombreuses mesures de bonne qualité de l'ajustement, comparables au  $R^2$  pour les modèles de régression, ont été proposées pour les modèles à réponse binaire, et de nombreuses applications statistiques reportent certaines d'entre elles. Consulter, parmi d'autres, Cragg et Uhler (1970), McFadden (1974a), Hauser (1977), Efron (1978), Amemiya (1981), et Maddala (1983). Le plus simple de ces pseudo  $R^2$  est celui suggéré par McFadden. Il est simplement défini comme

$$1 - \frac{\ell_U}{\ell_R}, \quad (15.18)$$

où  $\ell_U$  est la valeur non contrainte  $\ell(\mathbf{y}, \hat{\beta}_1, \hat{\beta}_2)$ , et  $\ell_R$  est la valeur contrainte  $\ell(\mathbf{y}, \bar{\beta}_1, \mathbf{0})$ . L'expression (15.18) représente une possible mesure de bonne qualité de l'ajustement parce qu'elle doit être comprise entre 0 et 1. Nous avons vu auparavant que la fonction de logvraisemblance (15.09) pour les modèles à choix binaires est bornée supérieurement par 0, ce qui implique que  $\ell_U$  et  $\ell_R$  sont toujours de même signe à moins que  $\ell_U$  soit nulle. Mais  $\ell_U$  peut être nulle seulement si le modèle non contraint s'ajuste parfaitement, ce qui survient s'il existe une classifcatrice parfaite. Ainsi nous voyons que l'expression (15.18) sera égale à 1 dans ce cas, égale à 0 quand les valeurs contrainte et non contrainte de la logvraisemblance seront identiques, et comprise entre 0 et 1 dans tous les autres cas.

Bien que (15.18) et d'autres mesures de bonne qualité d'ajustement puissent être utiles pour obtenir une idée approximative sur les performances d'un modèle à réponse binaire particulier, il n'est pas nécessaire de les utiliser si l'objectif est de comparer la performance de deux ou plusieurs modèles à réponse binaire différents estimés sur le même ensemble de données. Le meilleur moyen d'y parvenir consiste simplement à comparer les valeurs des fonctions de logvraisemblance, en utilisant le fait que les valeurs pour n'importe quel modèle à réponse binaire de la forme (15.03) sont directement comparables. Parfois, nous pouvons même rejeter un modèle sur la base d'une telle comparaison. Par exemple, supposons que, sur un ensemble de données particulier, la valeur de la logvraisemblance pour un modèle logit donné excède de plus de 1.92 celle d'un modèle probit avec la même fonction indice, ce qui représente la moitié de 3.84, la valeur critique à 5% pour une statistique de test qui est distribuée suivant une  $\chi^2(1)$ . Il est clairement possible d'englober les logit et probit dans un modèle plus général ayant plus d'un paramètre. Le dernier modèle s'ajusterait au moins aussi bien que le modèle logit; consulter la discussion dans la Section 14.3. Ainsi, dans cet exemple, nous pourrions rejeter à un niveau de 5% l'hypothèse selon laquelle le modèle probit a généré les observations. Naturellement, il est rare que la différence entre l'ajustement des modèles probit et logit, qui ne diffèrent d'aucune autre manière, soit aussi importante, à moins que la taille de l'échantillon ne soit extrêmement grande.



## 15.4 UNE RÉGRESSION ARTIFICIELLE

Il existe une régression à la fois très simple et très utile pour les modèles à réponse binaire. Comme pour d'autres régressions artificielles, elle peut être utilisée pour une variété d'usages, incluant l'estimation paramétrique, l'estimation de la matrice de covariance, et le test d'hypothèse. Cette régression artificielle a été suggérée par Engle (1984) et Davidson et MacKinnon (1984b). Elle peut être dérivée de plusieurs manières, parmi lesquelles la plus facile consiste à la traiter comme une version modifiée de la régression de Gauss-Newton.

Comme nous l'avons vu, le modèle à réponse binaire (15.03) peut être écrit sous la forme du modèle de régression non linéaire (15.11), soit  $y_t = F(\mathbf{X}_t\boldsymbol{\beta}) + e_t$ . Nous avons également vu que l'aléa  $e_t$  est de variance

$$V(\mathbf{X}_t\boldsymbol{\beta}) \equiv F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta})), \quad (15.19)$$

qui implique que (15.11) doit être estimée par GNLS. La GNR ordinaire correspondant à (15.11) serait

$$y_t - F(\mathbf{X}_t\boldsymbol{\beta}) = f(\mathbf{X}_t\boldsymbol{\beta})\mathbf{X}_t\mathbf{b} + \text{résidu}, \quad (15.20)$$

mais celle-ci est clairement inappropriée en raison de l'hétéroscédasticité des  $e_t$ . En effet, nous devons multiplier les deux membres de (15.20) par la racine carrée de l'inverse de (15.19). Ceci fournit la régression artificielle

$$(V(\mathbf{X}_t\boldsymbol{\beta}))^{-1/2}(y_t - F(\mathbf{X}_t\boldsymbol{\beta})) = (V(\mathbf{X}_t\boldsymbol{\beta}))^{-1/2}f(\mathbf{X}_t\boldsymbol{\beta})\mathbf{X}_t\mathbf{b} + \text{résidu}, \quad (15.21)$$

qui ressemble à la GNR pour un modèle de régression non linéaire estimé par moindres carrés pondérés (consulter la Section 9.4). La régression (15.21) est un cas particulier de ce que nous appellerons **régression pour modèle à réponse binaire**, ou **BRMR**. Cette forme de la BRMR demeure valable pour n'importe quel modèle à réponse binaire de la forme de (15.03).<sup>4</sup> Dans le cas du modèle logit, celle-ci se simplifie en

$$(\lambda(\mathbf{X}_t\boldsymbol{\beta}))^{-1/2}(y_t - \Lambda(\mathbf{X}_t\boldsymbol{\beta})) = (\lambda(\mathbf{X}_t\boldsymbol{\beta}))^{1/2}\mathbf{X}_t\mathbf{b} + \text{résidu}.$$

La BRMR satisfait les propriétés générales des régressions artificielles dont nous avons discuté dans la Section 14.4. En particulier, celle-ci est très

<sup>4</sup> Certains auteurs écrivent la BRMR de manières quelque peu différentes. Par exemple, chez Davidson et MacKinnon (1984b), la régressande a été définie comme

$$y_t \left( \frac{1 - F(\mathbf{X}_t\boldsymbol{\beta})}{F(\mathbf{X}_t\boldsymbol{\beta})} \right)^{1/2} + (y_t - 1) \left( \frac{F(\mathbf{X}_t\boldsymbol{\beta})}{1 - F(\mathbf{X}_t\boldsymbol{\beta})} \right)^{1/2}.$$

Vérifier qu'il s'agit juste d'une autre manière d'écrire la régressande de (15.21) constitue un bon exercice.

étroitement reliée à la fois au gradient de la fonction de vraisemblance (15.09) et à la matrice d'information. Le produit de la transposée de la régressande par la matrice des régresseurs fournit un vecteur d'élément type

$$\sum_{t=1}^n \frac{(y_t - F(\mathbf{X}_t\boldsymbol{\beta}))f(\mathbf{X}_t\boldsymbol{\beta})X_{ti}}{F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta}))},$$

qui est un élément type du vecteur gradient pour la fonction de logvraisemblance (15.09). La transposée de la matrice des régresseurs multipliée par elle-même fournit une matrice d'élément type

$$\sum_{t=1}^n \frac{f^2(\mathbf{X}_t\boldsymbol{\beta})}{F(\mathbf{X}_t\boldsymbol{\beta})(1 - F(\mathbf{X}_t\boldsymbol{\beta}))} X_{ti}X_{tj}. \quad (15.22)$$

La limite en probabilité de  $n^{-1}$  fois (15.22) est un élément type de la matrice d'information (15.13).

Toutes les fois que la fonction de logvraisemblance est globalement concave, comme pour les modèles logit et probit, il existe de nombreuses manières différentes d'estimer facilement les modèles à réponse binaire. Une approche qui fonctionne généralement bien consiste à utiliser un algorithme similaire à ceux décrits dans la Section 6.8. Dans un tel algorithme, la BRMR est utilisée pour déterminer la direction dans laquelle  $\boldsymbol{\beta}$  varie à chaque étape. Les valeurs de  $\boldsymbol{\beta}$  aux itérations  $j+1$  et  $j$  sont reliées par

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + \alpha^{(j)}\mathbf{b}^{(j)},$$

où  $\mathbf{b}^{(j)}$  désigne le vecteur des estimations OLS à partir de la BRMR (15.21) évaluée en  $\boldsymbol{\beta}^{(j)}$ , et  $\alpha^{(j)}$  est un scalaire déterminé par l'algorithme. On pourrait choisir les estimations initiales  $\boldsymbol{\beta}^{(1)}$  de différentes façons. Une de ces façons facile à utiliser et qui semble bien fonctionner dans la pratique consiste simplement à initialiser le terme constant à  $F^{-1}(\bar{y})$  et les autres coefficients à zéro. Les valeurs de départ correspondent alors aux estimations du modèle contraint avec des pentes nulles.

En évaluant la BRMR avec les estimations ML  $\hat{\boldsymbol{\beta}}$ , celle-ci peut aussi être utilisée pour obtenir une matrice de covariance estimée pour les paramètres estimés. La matrice de covariance estimée à partir de l'estimation OLS de la régression (15.21) évaluée en  $\hat{\boldsymbol{\beta}}$  sera

$$s^2(\mathbf{X}^\top \hat{\boldsymbol{\Psi}} \mathbf{X})^{-1}, \quad (15.23)$$

où  $s$  est l'écart type de la régression. Cet écart type tendra asymptotiquement vers 1, mais il ne sera pas vraiment égal à 1 dans les échantillons finis. La matrice  $\hat{\boldsymbol{\Psi}}$  est une matrice diagonale avec comme élément type diagonal

$$\hat{\Psi}_{tt} = \frac{f^2(\mathbf{X}_t\hat{\boldsymbol{\beta}})}{F(\mathbf{X}_t\hat{\boldsymbol{\beta}})(1 - F(\mathbf{X}_t\hat{\boldsymbol{\beta}}))}.$$

Il s'agit simplement de l'expression (15.12) avec  $\beta$  remplacé par  $\hat{\beta}$ . Ainsi, la matrice de covariance OLS estimée (15.23) fournit une estimation valable de la matrice de covariance de  $\hat{\beta}$ . C'est aussi le cas de la matrice  $(\mathbf{X}^\top \hat{\Psi} \mathbf{X})^{-1}$ , qui correspond simplement à (15.23) divisée par  $s^2$ , et que l'on préférera probablement utiliser puisque le facteur de  $s^2$  dans (15.23) introduit simplement un aléa additionnel dans l'estimation de la matrice de covariance.

Comme d'habitude, nous pouvons également estimer la matrice de covariance de  $\hat{\beta}$  par l'opposée de l'inverse de la matrice Hessienne numérique ou par le produit extérieur du gradient de la matrice CG,  $\hat{\mathbf{G}}^\top \hat{\mathbf{G}}$ . Dans le cas du modèle logit, l'opposée de la matrice Hessienne numérique est véritablement égale à la matrice d'information estimée  $\mathbf{X}^\top \hat{\Psi} \mathbf{X}$ , parce que

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_i \partial \beta_j} = \frac{\partial}{\partial \beta_j} \left( \sum_{t=1}^n (y_t - \Lambda(\mathbf{X}_t \beta)) X_{ti} \right) = - \sum_{t=1}^n \lambda(\mathbf{X}_t \beta) X_{ti} X_{tj}.$$

Cependant, dans le cas de la plupart des modèles à réponse binaire, incluant le modèle probit, l'opposée de la matrice Hessienne différera et sera généralement plus compliquée que la matrice d'information.

Comme toutes les régressions artificielles, la BRMR est particulièrement utile pour les tests d'hypothèse. Supposons que  $\beta$  soit partitionné comme  $[\beta_1 : \beta_2]$ , où  $\beta_1$  est un vecteur de dimension  $(k - r)$  et  $\beta_2$  est un vecteur de dimension  $r$ . Si  $\tilde{\beta}$  désigne le vecteur des estimations ML soumises à la contrainte  $\beta_2 = \mathbf{0}$ , nous pouvons tester cette contrainte en exécutant à la BRMR

$$\tilde{V}_t^{-1/2} (y_t - \tilde{F}_t) = \tilde{V}_t^{-1/2} \tilde{f}_t \mathbf{X}_{t1} \mathbf{b}_1 + \tilde{V}_t^{-1/2} \tilde{f}_t \mathbf{X}_{t2} \mathbf{b}_2 + \text{résidu}, \quad (15.24)$$

où  $\tilde{F}_t \equiv F(\mathbf{X}_t \tilde{\beta})$ ,  $\tilde{f}_t \equiv f(\mathbf{X}_t \tilde{\beta})$ , et  $\tilde{V}_t \equiv V(\mathbf{X}_t \tilde{\beta})$ . Ici  $\mathbf{X}_t$  a été partitionnée en deux vecteurs,  $\mathbf{X}_{t1}$  et  $\mathbf{X}_{t2}$ , correspondant à la partition de  $\beta$ . Les régresseurs qui correspondent à  $\beta_1$  sont orthogonaux à la régressande, tandis que ceux qui correspondent à  $\beta_2$  ne le sont pas. Toutes les statistiques de test usuelles pour  $\mathbf{b}_2 = \mathbf{0}$  sont valables. Cependant, par contraste avec le cas de la régression de Gauss-Newton, il n'existe pas de raison particulière d'utiliser un test en  $F$ , parce qu'il n'y a pas de paramètre de variance à estimer. La meilleure statistique de test à utiliser en échantillons finis, selon les résultats Monte Carlo obtenus par Davidson et MacKinnon (1984b), est probablement la somme des carrés expliqués à partir de la régression (15.24). Elle sera asymptotiquement distribuée suivant une  $\chi^2(r)$  sous l'hypothèse nulle. Notons que le  $nR^2$  ne sera pas égal à la somme des carrés expliqués dans ce cas, parce que la somme des carrés totaux ne sera pas égale à  $n$ .

Dans un cas très spécial, la BRMR (15.24) devient extrêmement simple. Supposons que l'hypothèse nulle corresponde à la nullité de tous les coefficients de pentes. Dans ce cas,  $\mathbf{X}_{t1}$  est unitaire,  $\mathbf{X}_t \tilde{\beta} = \tilde{\beta}_1 = F^{-1}(\bar{y})$ , et, dans une notation évidente, la régression (15.24) devient

$$\bar{V}^{-1/2} (y_t - \bar{F}) = \bar{V}^{-1/2} \bar{f} b_1 + \bar{V}^{-1/2} \bar{f} \mathbf{X}_{t2} \mathbf{b}_2 + \text{résidu}.$$

La statistique de test en  $F$  pour  $\mathbf{b}_2 = \mathbf{0}$  est invariante à la soustraction d'une constante à la régressande, ou à la multiplication de la régressande et des régresseurs par une constante. Ainsi, il est clair que nous pouvons tester l'hypothèse que toutes les pentes sont nulles en calculant simplement une statistique en  $F$  pour  $\mathbf{c}_2 = \mathbf{0}$  dans la régression linéaire

$$\mathbf{y} = c_1 + \mathbf{X}_2 \mathbf{c}_2 + \text{résidus}.$$

Ainsi, nous avons rencontré une situation dans laquelle le modèle de probabilité linéaire est utile. Si nous voulons tester l'hypothèse nulle selon laquelle aucun des régresseurs n'explique la variation de la variable dépendante, alors il est parfaitement pertinent d'employer la statistique de test ordinaire en  $F$  pour toutes les pentes nulles dans une régression OLS de  $\mathbf{y}$  sur  $\mathbf{X}$ .

Naturellement, nous pouvons utiliser la BRMR pour calculer les tests  $C(\alpha)$  et les tests pseudo-Wald aussi bien que des tests LM. L'essentiel de ce que nous avons dit concernant de tels tests dans les Sections 6.7 et 13.7 reste valable dans le contexte des modèles à réponse binaire. Nous ne pouvons pas utiliser la somme des carrés expliqués comme statistique de test, mais plutôt la réduction dans la somme des carrés expliqués consécutive à l'addition des régresseurs de test. Les tests pseudo-Wald peuvent être particulièrement utiles quand la fonction indice est linéaire sous l'hypothèse alternative mais non linéaire sous l'hypothèse nulle, parce que l'hypothèse alternative peut être estimée au moyen d'un programme standard logit ou probit. S'il apparaît que les contraintes s'ajustent bien aux données, nous pouvons employer une BRMR différente pour obtenir des estimations en une étape.

La BRMR est utile pour tester tous les aspects de la spécification des modèles à réponse binaire. Avant même d'accepter un quelconque modèle de la sorte, nous devons savoir si  $F(\mathbf{X}_t \boldsymbol{\beta})$  représente une spécification correcte pour la probabilité  $y_t = 1$  conditionnellement à l'ensemble d'information  $\Omega_t$ . Les tests de variables appartenant à l'ensemble  $\Omega_t$  potentiellement omises constitue une part importante de ce processus, et nous avons déjà vu comment procéder à l'aide de la BRMR (15.24). Mais même si  $\mathbf{X}_t$  est spécifiée de façon correcte, le reste du modèle peut ne pas l'être.

Considérons le modèle à variable latente donné par (15.05) et (15.06). Parce que les modèles à réponse binaire sont typiquement estimés en utilisant les données en coupe transversale, et que de telles données présentent fréquemment de l'hétéroscédasticité, il est fort possible que les aléas dans l'équation pour  $y_t^*$  soient hétéroscédastiques. S'ils étaient effectivement hétéroscédastiques, le modèle probit ne serait plus approprié, et les estimations de  $\boldsymbol{\beta}$  basées sur ce modèle seraient non convergentes; consulter Yatchew et Griliches (1984). Puisque nous pouvons considérer que tout modèle à réponse binaire peut provenir d'un modèle à variable latente, il est clairement important de tester l'hétéroscédasticité de tels modèles. Nous discutons à présent de la manière de procéder.

Une spécification plus générale que l'équation (15.05) qui tient compte des erreurs hétéroscédastiques est

$$y_t^* = \mathbf{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim N(0, \exp(2\mathbf{Z}_t\boldsymbol{\gamma})), \quad (15.25)$$

où  $\mathbf{Z}_t$  est un vecteur ligne de longueur  $q$  des observations sur les variables qui appartiennent à l'ensemble d'information  $\Omega_t$ . Pour s'assurer qu'à la fois  $\boldsymbol{\beta}$  et  $\boldsymbol{\gamma}$  sont identifiables,  $\mathbf{Z}_t$  ne doit pas comprendre un terme constant ou l'équivalent. La combinaison de (15.25) avec (15.06) fournit le modèle

$$E(y_t | \Omega_t) = \Phi\left(\frac{\mathbf{X}_t\boldsymbol{\beta}}{\exp(\mathbf{Z}_t\boldsymbol{\gamma})}\right). \quad (15.26)$$

Quand  $\boldsymbol{\gamma} = \mathbf{0}$ , (15.25) se réduit à (15.05) et (15.26) se réduit au modèle probit ordinaire. Même quand un modèle à réponse binaire autre que le modèle probit est utilisé, il semble encore très raisonnable de considérer l'hypothèse alternative

$$E(y_t | \Omega_t) = F\left(\frac{\mathbf{X}_t\boldsymbol{\beta}}{\exp(\mathbf{Z}_t\boldsymbol{\gamma})}\right).$$

Nous pouvons tester  $\boldsymbol{\gamma} = \mathbf{0}$  contre cette forme d'hétéroscédasticité. La BRMR appropriée est

$$\hat{V}_t^{-1/2}(y_t - \hat{F}_t) = \hat{V}_t^{-1/2}\hat{f}_t\mathbf{X}_t\mathbf{b} + \hat{V}_t^{-1/2}\hat{f}_t\mathbf{Z}_t(-\mathbf{X}_t\hat{\boldsymbol{\beta}})\mathbf{c} + \text{résidu}, \quad (15.27)$$

où  $\hat{F}_t$ ,  $\hat{f}_t$ , et  $\hat{V}_t$  sont évalués avec les estimations ML  $\hat{\boldsymbol{\beta}}$  en supposant que  $\boldsymbol{\gamma} = \mathbf{0}$ . La somme expliquée des carrés de (15.27) sera distribuée asymptotiquement suivant une  $\chi^2(q)$  sous l'hypothèse nulle.

Il est également important de tester la spécification de la fonction de transformation  $F(\cdot)$ . Comme nous l'avons noté plus tôt, une manière naturelle de procéder de la sorte consiste à considérer un modèle alternatif de la forme

$$E(y_t | \Omega_t) = F(h(\mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\alpha})), \quad (15.28)$$

où  $h(x, \boldsymbol{\alpha})$  est une fonction non linéaire de  $x$ , et  $\boldsymbol{\alpha}$  est soit un paramètre soit un vecteur de paramètres tel que  $h(\mathbf{X}_t\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{X}_t\boldsymbol{\beta}$  pour une certaine valeur de  $\boldsymbol{\alpha}$ . Stukel (1988) suggère une famille plus compliquée de fonctions à deux paramètres  $h(x, \boldsymbol{\alpha})$  qui mène à une famille très générale de modèles. Cette famille comprend le modèle logit comme un cas particulier quand  $\boldsymbol{\alpha} = \mathbf{0}$ , et permet d'imposer ou non l'hypothèse de symétrie-oblique. On peut aisément utiliser la BRMR pour tester l'hypothèse nulle que  $\boldsymbol{\alpha} = \mathbf{0}$  contre cette alternative.

Un test plus simple peut être basé sur la famille de modèles

$$E(y_t | \Omega_t) = F\left(\frac{\tau(\alpha\mathbf{X}_t\boldsymbol{\beta})}{\alpha}\right),$$

qui est un cas particulier de (15.28). Ici  $\tau(\cdot)$  peut être n'importe quelle fonction monotone croissante en son argument satisfaisant les conditions

$$\tau(0) = 0, \quad \tau'(0) = 1, \quad \text{et} \quad \tau''(0) \neq 0.$$

En utilisant la Règle de l'Hôpital, MacKinnon et Magee (1990) montrent que

$$\lim_{\alpha \rightarrow 0} \left( \frac{\tau(\alpha x)}{\alpha} \right) = x \quad \text{et} \quad \lim_{\alpha \rightarrow 0} \left( \frac{\partial(\tau(\alpha x)/\alpha)}{\partial \alpha} \right) = \frac{1}{2} x^2 \tau''(0). \quad (15.29)$$

Ainsi pour tester l'hypothèse nulle que  $\alpha = 0$ , la BRMR est

$$\hat{V}_t^{-1/2}(y_t - \hat{F}_t) = \hat{V}_t^{-1/2} \hat{f}_t \mathbf{X}_t \mathbf{b} + a \hat{V}_t^{-1/2} (\mathbf{X}_t \hat{\boldsymbol{\beta}})^2 \hat{f}_t + \text{résidu}, \quad (15.30)$$

où le terme constant  $\tau''(0)/2$  qui provient de (15.29) est non pertinent pour le test et a été omis. Par conséquent, la régression (15.30) traite simplement les valeurs au carré de la fonction indice évaluée en  $\hat{\boldsymbol{\beta}}$  comme s'il y avait des observations sur un régresseur potentiellement omis. Ce test comporte une forte ressemblance avec le test RESET pour les modèles de régression, dont nous avons discuté dans la Section 6.5. Nous pouvons utiliser comme statistique de test la statistique ordinaire  $t$  pour  $a = 0$ , mais nous préférons employer la somme expliquée des carrés.

Une très grande variété de tests de spécification peut être basée sur la BRMR. En réalité, presque tous les tests de spécification pour les modèles de régression basés sur une régression artificielle comportent un analogue pour les modèles à réponse binaire. En général, nous pouvons écrire la régression artificielle pour la réalisation d'un tel test comme

$$\hat{V}_t^{-1/2}(y_t - \hat{F}_t) = \hat{V}_t^{-1/2} \hat{f}_t \mathbf{X}_t \mathbf{b} + \hat{\mathbf{Z}}_t \mathbf{c} + \text{résidu}, \quad (15.31)$$

où  $\hat{\mathbf{Z}}_t$  est un vecteur de dimension  $1 \times r$  qui peut dépendre des estimations ML  $\hat{\boldsymbol{\beta}}$  et d'un élément quelconque dans l'ensemble d'information  $\Omega_t$ . L'intuition pour (15.31) est très simple. Si  $F(\mathbf{X}_t \boldsymbol{\beta})$  est la spécification correcte de  $E(y_t | \Omega_t)$ , alors (15.31) *sans* les régresseurs  $\mathbf{Z}_t$  est la régression artificielle qui correspond au DGP. Aucun régresseurs additionnels  $\mathbf{Z}_t$  dépendant de  $\Omega_t$  ne doit avoir de pouvoir explicatif significatif lorsqu'il est ajouté à cette régression.

Il est même possible d'utiliser la BRMR pour calculer les tests non emboîtés très similaires au test en  $P$  (consulter la Section 11.3). Supposons que nous avons deux modèles en compétition:

$$H_1: E(y_t | \Omega_t) = F_1(\mathbf{X}_{1t} \boldsymbol{\beta}_1) \quad \text{et}$$

$$H_2: E(y_t | \Omega_t) = F_2(\mathbf{X}_{2t} \boldsymbol{\beta}_2),$$

qui peuvent différer soit parce que  $F_1(\cdot)$  n'est pas la même que  $F_2(\cdot)$  soit parce que  $\mathbf{X}_{1t}$  n'est pas la même que  $\mathbf{X}_{2t}$  soit pour les deux raisons à la

fois. Il existe de nombreuses manières d'emboîter  $H_1$  et  $H_2$  dans un modèle composite artificiel. Une des plus simples est

$$H_C: E(y_t | \Omega_t) = (1 - \alpha)F_1(\mathbf{X}_{1t}\boldsymbol{\beta}_1) + \alpha F_2(\mathbf{X}_{2t}\boldsymbol{\beta}_2),$$

bien que ce modèle artificiel ne soit pas vraiment un modèle à réponse binaire. Nous pouvons tester  $H_1$  contre  $H_C$  essentiellement de la manière dont nous l'avons fait pour les modèles de régression. Tout d'abord, nous remplaçons  $\boldsymbol{\beta}_2$  par son estimation ML  $\hat{\boldsymbol{\beta}}_2$  et construisons ensuite une régression artificielle pour tester l'hypothèse nulle que  $\alpha = 0$ . Cette régression artificielle est

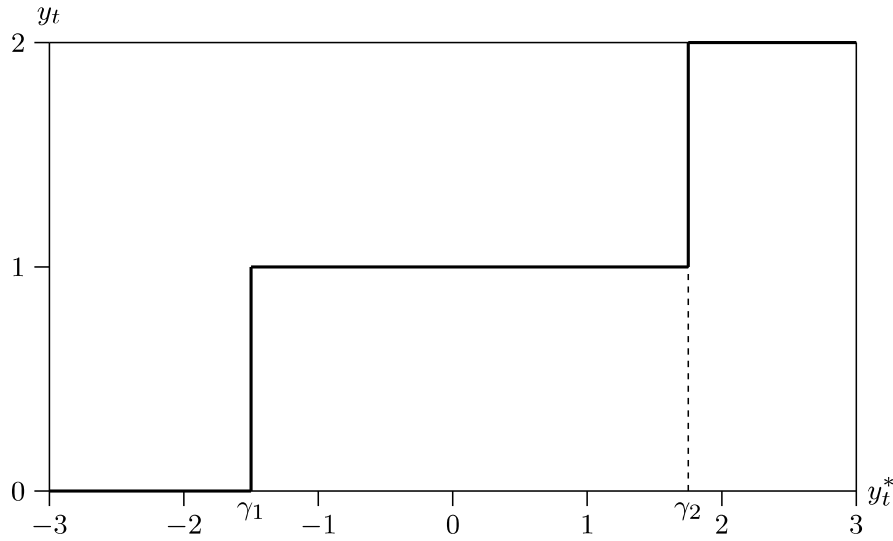
$$\hat{V}_t^{-1/2}(y_t - \hat{F}_{1t}) = \hat{V}_t^{-1/2}\hat{f}_{1t}\mathbf{X}_{1t}\mathbf{b} + a\hat{V}_t^{-1/2}(\hat{F}_{2t} - \hat{F}_{1t}) + \text{résidu}.$$

Le régresseur de test est simplement la différence entre les probabilités que  $y_t = 1$  selon les deux modèles, multipliée par  $\hat{V}_t^{-1/2}$ , le facteur de pondération qui est aussi utilisé pour la régressande et pour les autres régresseurs.

L'estimation des deux modèles standards à réponse binaire, c'est-à-dire les modèles probit et logit avec des fonctions indice linéaires, est extrêmement facile avec la plupart des applications de régression, et l'estimation des modèles qui impliquent des fonctions de transformation non standards et/ou des fonctions indice non linéaires n'est généralement pas très difficile. Comme les tests de tels modèles au moyen de la BRMR sont également très faciles, il n'existe absolument aucune excuse pour tester de façon moins complète les spécifications des modèles à réponse binaire que celles des modèles de régression.

## 15.5 LES MODÈLES À PLUS DE DEUX RÉPONSES DISCRÈTES

Bien que de nombreuses variables dépendantes discrètes soient binaires, des variables discrètes qui peuvent prendre trois ou plusieurs valeurs différentes sont assez fréquentes en économie. Une variété de modèles à réponse qualitative a été inventée pour traiter de tels cas. Ceux-ci se répartissent en deux catégories: les modèles conçus pour traiter des **réponses ordonnées** et ceux conçus pour traiter des **réponses non ordonnées**. Un exemple de données à réponse ordonnée serait les résultats provenant d'une enquête où il a été demandé aux personnes interrogées de dire si elles sont entièrement d'accord, d'accord, sans opinion, en désaccord, ou entièrement en désaccord avec un quelconque sujet. Ici, il y a cinq réponses possibles, qui peuvent bien évidemment être ordonnées de manière graduée. Un exemple de données à réponse non ordonnée serait les résultats provenant d'une enquête sur le moyen de transport que choisissent les gens pour faire la navette entre leur résidence et leur travail. Les réponses possibles peuvent être: marche, vélo, prendre le bus, usage commun d'un véhicule, et usage individuel d'un véhicule. Bien qu'il puisse exister plusieurs manières d'ordonner ces réponses, aucune ne s'impose naturellement.



**Figure 15.3** Relation entre  $y_t^*$  et  $y_t$  dans un modèle probit ordonné

La manière la plus commune de traiter les données à réponse ordonnée consiste à utiliser un **modèle à réponse qualitative ordonnée**, habituellement soit le **modèle probit ordonné** soit le **modèle logit ordonné**. Comme exemple, considérons le modèle à variable latente

$$y_t^* = \mathbf{X}_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, 1), \quad (15.32)$$

où, pour une raison qui deviendra bientôt évidente,  $\mathbf{X}_t$  ne comprend aucun terme constant. Nous observons réellement une variable discrète  $y_t$  qui peut prendre seulement trois valeurs:

$$\begin{aligned} y_t &= 0 \quad \text{si } y_t^* < \gamma_1 \\ y_t &= 1 \quad \text{si } \gamma_1 \leq y_t^* < \gamma_2 \\ y_t &= 2 \quad \text{si } \gamma_2 \leq y_t^*. \end{aligned} \quad (15.33)$$

Les paramètres de ce modèle sont  $\boldsymbol{\beta}$  et  $\boldsymbol{\gamma} \equiv [\gamma_1 : \gamma_2]$ . Les  $\gamma_i$  sont des seuils qui déterminent quelle valeur de  $y_t$  va correspondre à une valeur donnée de  $y_t^*$ . Ceci est illustré dans la Figure 15.3. Le vecteur  $\boldsymbol{\gamma}$  comporte toujours un élément de moins qu'il y a de choix. Quand il existe seulement deux choix, ce modèle se confond avec le modèle à réponse binaire ordinaire, l'unique élément de  $\boldsymbol{\gamma}$  jouant le rôle du terme constant.

La probabilité que  $y_t = 0$  est

$$\begin{aligned} \Pr(y_t = 0) &= \Pr(y_t^* < \gamma_1) = \Pr(\mathbf{X}_t \boldsymbol{\beta} + u_t < \gamma_1) \\ &= \Pr(u_t < \gamma_1 - \mathbf{X}_t \boldsymbol{\beta}) \\ &= \Phi(\gamma_1 - \mathbf{X}_t \boldsymbol{\beta}). \end{aligned}$$



De façon similaire, la probabilité que  $y_t = 1$  est

$$\begin{aligned}\Pr(y_t = 1) &= \Pr(\gamma_1 \leq y_t^* < \gamma_2) = \Pr(\gamma_1 \leq \mathbf{X}_t\boldsymbol{\beta} + u_t < \gamma_2) \\ &= \Pr(u_t < \gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) - \Pr(u_t \leq \gamma_1 - \mathbf{X}_t\boldsymbol{\beta}) \\ &= \Phi(\gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) - \Phi(\gamma_1 - \mathbf{X}_t\boldsymbol{\beta}),\end{aligned}$$

et la probabilité que  $y_t = 2$  est

$$\begin{aligned}\Pr(y_t = 2) &= \Pr(y_t^* \geq \gamma_2) = \Pr(\mathbf{X}_t\boldsymbol{\beta} + u_t \geq \gamma_2) \\ &= \Pr(u_t \geq \gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) \\ &= \Phi(\mathbf{X}_t\boldsymbol{\beta} - \gamma_2).\end{aligned}$$

Ainsi, la fonction de logvraisemblance pour le modèle probit ordonné composé de (15.32) et (15.33) est

$$\begin{aligned}\ell(\boldsymbol{\beta}, \gamma_1, \gamma_2) &= \sum_{y_t=0} \log(\Phi(\gamma_1 - \mathbf{X}_t\boldsymbol{\beta})) + \sum_{y_t=2} \log(\Phi(\mathbf{X}_t\boldsymbol{\beta} - \gamma_2)) \\ &\quad + \sum_{y_t=1} \log(\Phi(\gamma_2 - \mathbf{X}_t\boldsymbol{\beta}) - \Phi(\gamma_1 - \mathbf{X}_t\boldsymbol{\beta})).\end{aligned}\tag{15.34}$$

Notons que  $\gamma_2$  doit être plus grand que  $\gamma_1$ , parce que sinon  $\Pr(y_t = 1)$  serait négative et le dernier terme dans (15.34) serait indéfini.

La maximisation de la fonction de logvraisemblance (15.34) est relativement simple, comme l'est généralement le modèle qui gère plus de trois réponses. Il est aussi évident que l'on pourrait utiliser une quelque autre fonction de transformation à la place de la normale standard dans (15.34) et avoir encore un modèle parfaitement correct. La forme de la fonction de logvraisemblance serait malgré tout inchangée. Pour une discussion plus approfondie concernant les modèles à réponse qualitative, consulter Greene (1990a, Chapitre 20), Terza (1985), et Becker et Kennedy (1992). Cette approche n'est en aucune manière la seule pour traiter des réponses discrètes ordonnées. D'autres approches sont examinées par McCullagh (1980), Agresti (1984), et Rahiala et Teräsvirta (1988).

La caractéristique majeure des modèles à réponse qualitative ordonnée est que tous les choix dépendent d'une seule fonction indice. Ceci prend tout son sens quand les réponses ont un ordre naturel mais pas dans le cas contraire. Un type de modèle différent est évidemment nécessaire pour traiter des réponses non ordonnées. L'approche la plus simple consiste à employer le **modèle logit multinomial** (ou **modèle logit multiple**), qui a été largement utilisé dans des travaux appliqués. Un premier exemple est Schmidt et Strauss (1975). Un modèle relativement proche appelé **modèle logit conditionnel** est aussi largement utilisé, comme nous le verrons plus tard.<sup>5</sup>

<sup>5</sup> La terminologie dans ce domaine est souvent utilisée de différentes manières suivant les auteurs. Les termes "modèle logit multinomial," "modèle logit multiple," et "modèle logit conditionnel" sont parfois utilisés de façon interchangeable.

Le modèle logit multinomial est conçu pour traiter  $J + 1$  réponses. Selon ce modèle, la probabilité d'observer chacune d'entre elles est

$$\Pr(y_t = 0) = \frac{1}{1 + \sum_{j=1}^J \exp(\mathbf{X}_t \boldsymbol{\beta}^j)} \quad (15.35)$$

$$\Pr(y_t = l) = \frac{\exp(\mathbf{X}_t \boldsymbol{\beta}^l)}{1 + \sum_{j=1}^J \exp(\mathbf{X}_t \boldsymbol{\beta}^j)} \quad \text{pour } l = 1, \dots, J. \quad (15.36)$$

Ici  $\mathbf{X}_t$  est un vecteur ligne de dimension  $k$  d'observations sur les variables qui appartiennent à l'ensemble d'information d'intérêt, et  $\boldsymbol{\beta}^1$  jusqu'à  $\boldsymbol{\beta}^J$  sont des vecteurs de dimension  $k$  des paramètres. Quand  $J = 1$ , nous voyons aisément que ce modèle se réduit au modèle logit ordinaire avec une seule fonction indice  $\mathbf{X}_t \boldsymbol{\beta}^1$ . Pour chaque alternative additionnelle, nous ajoutons au modèle une autre fonction indice et  $k$  paramètres.

Certains auteurs préfèrent écrire le modèle logit multinomial comme

$$\Pr(y_t = l) = \frac{\exp(\mathbf{X}_t \boldsymbol{\beta}^l)}{\sum_{j=0}^J \exp(\mathbf{X}_t \boldsymbol{\beta}^j)} \quad \text{pour } l = 0, \dots, J \quad (15.37)$$

en définissant un vecteur paramétrique supplémentaire  $\boldsymbol{\beta}^0$ , dont tous les éléments sont nuls. Cette manière d'écrire le modèle est plus compacte que (15.35) et (15.36) mais ne montre pas clairement que le modèle logit ordinaire est un cas particulier du modèle logit multinomial.

L'estimation du modèle logit multinomial est raisonnablement simple, parce que la fonction de logvraisemblance est globalement concave. Cette fonction de logvraisemblance peut être écrite comme

$$\ell(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_J) = \sum_{j=1}^J \sum_{y_t=j} \mathbf{X}_t \boldsymbol{\beta}^j - \sum_{t=1}^n \log \left( 1 + \sum_{j=1}^J \exp(\mathbf{X}_t \boldsymbol{\beta}^j) \right).$$

Cette fonction est la somme des contributions de chaque observation. Chaque contribution comporte deux termes: le premier est  $\mathbf{X}_t \boldsymbol{\beta}^j$ , où l'indice  $j$  correspond à  $y_t = j$  (ou zéro si  $j = 0$ ), et le second est l'opposé du logarithme du dénominateur qui apparaît dans (15.35) et (15.36).

Une propriété importante du modèle logit multinomial est que

$$\frac{\Pr(y_t = l)}{\Pr(y_t = j)} = \frac{\exp(\mathbf{X}_t \boldsymbol{\beta}^l)}{\exp(\mathbf{X}_t \boldsymbol{\beta}^j)} = \exp(\mathbf{X}_t (\boldsymbol{\beta}^l - \boldsymbol{\beta}^j)) \quad (15.38)$$

pour les deux réponses  $l$  et  $j$  (en incluant la réponse nulle si nous interprétons  $\boldsymbol{\beta}^0$  comme un vecteur composé de zéros). Ainsi les disparités entre deux réponses quelconques ne dépendent que de  $\mathbf{X}_t$  et des vecteurs paramétriques associés à ces deux réponses. Elles ne dépendent pas des

vecteurs paramétriques associés à n'importe laquelle des autres réponses. En fait, nous voyons à partir de (15.38) que le logarithme des disparités entre les réponses  $l$  et  $j$  est simplement  $\mathbf{X}_i\boldsymbol{\beta}^*$ , où  $\boldsymbol{\beta}^* \equiv (\boldsymbol{\beta}^l - \boldsymbol{\beta}^j)$ . Ainsi, conditionnellement au choix de  $j$  ou  $l$ , le choix entre les deux réponses est déterminé par un modèle logit ordinaire avec un vecteur de paramètres  $\boldsymbol{\beta}^*$ .

Le **modèle logit conditionnel** est très étroitement relié au modèle logit multinomial, et fut employé pour la première fois par McFadden (1974a, 1974b). Consulter Domencich et McFadden (1975), McFadden (1984), et Greene (1990a, Chapitre 20) pour des traitements détaillés. Le modèle logit conditionnel est conçu pour traiter des choix du consommateur parmi  $J$  (et non pas  $J+1$ ) alternatives discrètes, où une et seulement une des ces alternatives peut être choisie. Supposons que quand le  $i^{\text{ième}}$  consommateur choisit l'alternative  $j$ , il ou elle obtient l'utilité

$$U_{ij} = \mathbf{W}_{ij}\boldsymbol{\beta} + \varepsilon_{ij},$$

où  $\mathbf{W}_{ij}$  est un vecteur ligne des caractéristiques de l'alternative  $j$  telles qu'elles s'appliquent au consommateur  $i$ . Soit  $y_i$  le choix réalisé par le  $i^{\text{ième}}$  consommateur. Il est vraisemblable que  $y_i = l$  si  $U_{il}$  est au moins aussi important que  $U_{ij}$  pour tout  $j \neq l$ . Alors si les perturbations  $\varepsilon_{ij}$  pour  $j = 1, \dots, J$  sont indépendantes et identiquement distribuées selon la distribution de Weibull, nous pouvons montrer que

$$\Pr(y_i = l) = \frac{\exp(\mathbf{W}_{il}\boldsymbol{\beta})}{\sum_{j=1}^J \exp(\mathbf{W}_{ij}\boldsymbol{\beta})}. \quad (15.39)$$

Ceci ressemble fort à (15.37), et il est aisé de voir que la somme de probabilités doit être égale à l'unité.

Il existe deux différences majeures entre le modèle logit multinomial et le modèle logit conditionnel. Dans le premier, il existe un seul vecteur de variables indépendantes pour chaque observation, et il y a  $J$  vecteurs différents de paramètres. Dans le second, les valeurs des variables indépendantes varient en fonction des alternatives, mais il y a juste un seul vecteur paramétrique  $\boldsymbol{\beta}$ . Le modèle logit multinomial est une généralisation directe du modèle logit qui peut être utilisé pour traiter une quelconque situation comprenant trois ou plusieurs réponses qualitatives non ordonnées. Par contraste, le modèle logit conditionnel est spécifiquement conçu pour traiter des choix du consommateur parmi des alternatives discrètes basés sur les caractéristiques de ces alternatives.

Il peut exister un certain nombre de subtilités associées à la spécification et à l'interprétation des modèles logit conditionnels selon la nature des variables explicatives. Il n'y a pas suffisamment de place dans cet ouvrage pour traiter celles-ci convenablement, et par conséquent les lecteurs qui désirent estimer de tels modèles sont encouragés à consulter les références mentionnées

auparavant. Une propriété importante des modèles logit conditionnels est l'analogie de (15.38):

$$\frac{\Pr(y_t = l)}{\Pr(y_t = j)} = \frac{\exp(\mathbf{W}_{il}\boldsymbol{\beta})}{\exp(\mathbf{W}_{ij}\boldsymbol{\beta})}. \quad (15.40)$$

Cette propriété est appelée **indépendance par rapport aux alternatives non pertinentes**, ou **IIA**. Elle implique que la prise en compte d'une alternative supplémentaire par le modèle, ou que le changement des caractéristiques d'une alternative déjà comprise par le modèle, ne modifiera pas les probabilités des alternatives  $l$  et  $j$ .

La propriété IIA peut être extrêmement peu plausible dans certaines circonstances. Supposons qu'il y ait initialement deux alternatives pour voyager entre deux villes: la liaison aérienne par Air Monopole et la route. Supposons par ailleurs que la moitié des voyageurs choisissent l'avion et l'autre moitié la voiture. Alors Air Concurrent entre sur le marché et crée une troisième alternative. Si Air Concurrent offre un service identique à celui de Air Monopole, il doit gagner la même part de marché. Ainsi, selon la propriété IIA, un tiers des voyageurs doit prendre chaque compagnie aérienne et un tiers doit prendre la route. Par conséquent, l'automobile a perdu une part de marché identique à celle de Air Monopole depuis l'entrée de Air Concurrent! Ceci semble très peu plausible.<sup>6</sup> La conséquence de tout ceci est qu'un grand nombre d'articles a été consacré au problème du test de la propriété d'indépendance aux alternatives non pertinentes et à la découverte de modèles réalisables qui ne possèdent pas cette propriété. Voir, en particulier, Hausman et Wise (1978), Manski et McFadden (1981), Hausman et McFadden (1984), et McFadden (1987).

Ceci termine notre discussion des modèles à réponse qualitative. Des traitements plus détaillés sont présentés dans les articles de synthèse de Maddala (1983), McFadden (1984), Amemiya (1981; 1985, Chapitre 9), et Greene (1990a, Chapitre 20), entre autres. Dans les trois prochaines sections, nous nous focaliserons sur les variables dépendantes limitées.

## 15.6 LES MODÈLES POUR DONNÉES AVEC TRONCATURE

Les modèles à variable dépendante limitée sont conçus pour traiter des échantillons **tronqués** ou **censurés** d'une certaine manière. Ces deux termes sont facilement confondus. Un échantillon a été tronqué si certaines

<sup>6</sup> En effet, nous pourrions objecter que la guerre des tarifs entre Air Monopole et Air Concurrent inciterait certains automobilistes à prendre l'avion. Alors l'automobile perdrait effectivement des parts de marché. Mais si les deux compagnies aériennes offraient des prix plus bas, un ou plusieurs éléments des  $\mathbf{W}_{ij}$  associés à ces prix seraient modifiés. L'analyse précédente suppose que tous les  $\mathbf{W}_{ij}$  restent inchangés.

de ses observations qui devaient y être ont été systématiquement exclues de l'échantillon. Par exemple, un échantillon de ménages avec des revenus inférieurs à \$100,000 exclut nécessairement tous les ménages ayant des revenus supérieurs à ce niveau. Il ne s'agit pas d'un échantillon aléatoire de tous les ménages. Si la variable dépendante est le revenu, ou une série corrélée avec le revenu, les résultats qui utilisent l'échantillon tronqué pourraient être potentiellement très fallacieux.

De l'autre côté, un échantillon a été censuré si aucune observation n'a été systématiquement exclue, mais si une certaine information contenues par ces observations a été supprimée. Songeons au "censeur" qui lit le courrier des gens et occulte certaines parties de celui-ci. Les destinataires reçoivent encore leur courrier, mais des passages de celui-ci sont illisibles. Pour continuer sur ce premier exemple, supposons que les ménages avec tous les niveaux de revenu soient inclus dans l'échantillon, mais que pour ceux dont les revenus excèdent \$100,000, le montant reporté est toujours exactement \$100,000.<sup>7</sup> Dans ce cas, l'échantillon censuré est encore un échantillon aléatoire de tous les ménages, mais les valeurs reportées pour les ménages à hauts revenus ne sont pas les véritables valeurs. Nous pouvons assimiler les variables dépendantes discrètes à un type de censure encore plus prononcé. Par exemple, si nous nous contentions de classer les revenus des ménages dans des intervalles en dollars, la variable dépendante serait des réponses qualitatives ordonnées. Cependant, censurer à ce point n'est pas habituellement considéré comme une censure.

Les économètres ont imaginé un grand nombre de modèles pour traiter les données tronquées et censurées. L'espace dont nous disposons ne nous autorise à traiter que quelques uns des plus simples. Greene (1990a, Chapitre 21) fournit un excellent article de synthèse récent. D'autres articles intéressants de tout ou partie de ce domaine sont ceux de Dhrymes (1986), Maddala (1983, 1986), et Amemiya (1984; 1985, Chapitre 10). En supplément, une parution du *Journal of Econometrics* (Blundell, 1987) est consacrée au thème important des tests de spécification dans les modèles à variable dépendante limitée (et aussi à réponse qualitative).

Tout d'abord, nous nous consacrerons à la plus simple sorte de modèle à variable dépendante tronquée. Supposons que pour tout  $t$  (observé ou non) l'espérance de  $y_t$  conditionnelle à l'ensemble d'information  $\Omega_t$  soit donnée par une fonction de régression non linéaire  $x_t(\beta)$ , qui pourrait bien être la fonction de régression linéaire  $\mathbf{X}_t\beta$ . Alors, si les aléas sont normalement et indépendamment distribués, nous pouvons écrire

$$y_t = x_t(\beta) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \quad (15.41)$$

<sup>7</sup> Ce type de censure est commun avec les données d'enquête. Il peut survenir soit parce que les enquêteurs désirent protéger l'intimité des personnes à haut revenu interrogées soit parce que l'enquête n'était pas conçue pour les nécessités de l'analyse économétrique que l'on souhaite mettre en oeuvre.

Supposons maintenant que  $y_t$  soit observée seulement si  $y_t \geq y^l$ , où  $y^l$  est une certaine borne inférieure fixée. La probabilité que  $y_t$  soit observée est

$$\begin{aligned}\Pr(y_t \geq y^l) &= \Pr(x_t(\beta) + u_t \geq y^l) = 1 - \Pr(u_t < y^l - x_t(\beta)) \\ &= 1 - \Phi\left(\frac{1}{\sigma}(y^l - x_t(\beta))\right) = \Phi\left(-\frac{1}{\sigma}(y^l - x_t(\beta))\right).\end{aligned}$$

Ainsi, quand  $x_t(\beta) = y^l$ , la probabilité que n'importe quelle observation soit observée est un demi. Comme  $x_t(\beta)$  croît (ou décroît) relativement à  $y^l$ , la probabilité d'observer aussi  $y_t$  augmente (ou diminue). Ceci constitue un exemple simple d'un **modèle de régression tronqué**.

La troncature peut être un problème pour l'estimation de (15.41) selon l'objectif de cette estimation. L'estimation par moindres carrés serait appropriée si nous étions intéressés par l'espérance de  $y_t$  conditionnelle à  $\Omega_t$  et conditionnelle à  $y_t$  pour des valeurs supérieures à  $y^l$ . Mais il est peu probable que cela nous intéresse. Nous avons défini  $x_t(\beta)$  comme l'espérance de  $y_t$  conditionnelle à  $\Omega_t$  sans aucune référence à  $y^l$ . Si c'est véritablement ce qui nous intéresse, les estimations par moindres carrés de (15.41) pourraient être sérieusement trompeuses.

Le problème est que l'espérance de  $u_t$  conditionnelle à  $y_t \geq y^l$  n'est pas nulle. L'élément  $y_t$  ne sera supérieur à  $y^l$  que si  $u_t$  est suffisamment grand, et alors seulement l'observation  $t$  fera partie de l'échantillon. Ainsi, pour des observations qui sont dans l'échantillon,  $E(u_t) > 0$ . De fait, nous pouvons montrer que

$$E(u_t | y_t \geq y^l) = \frac{\sigma \phi((y^l - x_t(\beta))/\sigma)}{\Phi(-(y^l - x_t(\beta))/\sigma)}. \quad (15.42)$$

Evidemment, l'espérance conditionnelle de  $u_t$  dans ce cas est positive et dépend de  $x_t(\beta)$ . Ce résultat utilise la propriété que si une variable aléatoire  $z$  est normale centrée et réduite, l'espérance de  $z$  conditionnelle à  $z \geq z^*$  est  $\phi(z^*)/\Phi(-z^*)$ ; consulter Johnson et Kotz (1970a). De façon similaire, l'espérance de  $z$  conditionnelle à  $z \leq z^*$  est  $-\phi(z^*)/\Phi(z^*)$ . Ainsi, si la troncature s'appliquait aux valeurs supérieures plutôt qu'inférieures, l'espérance conditionnelle de  $u_t$  serait négative au lieu d'être positive.

Nous ne pouvons pas raisonnablement obtenir des estimations convergentes de  $\beta$  en utilisant les estimations par moindres carrés quand les aléas ont une espérance positive (15.42) qui dépend de  $x_t(\beta)$ . Goldberger (1981) fournit quelques expressions pour l'importance de la non convergence dans certains cas, et dans la prochaine section (consulter le Tableau 15.1), nous fournissons quelques résultats numériques qui illustrent cette à quel point elle peut être conséquente. Le remède évident consiste à utiliser la méthode du maximum de vraisemblance. La densité de  $y_t$  conditionnelle à  $y_t \geq y^l$  est simplement la densité non conditionnelle de  $y_t$  restreinte aux valeurs de  $y_t \geq y^l$ , divisée par la probabilité que  $y_t \geq y^l$ :

$$\frac{\sigma^{-1} \phi((y_t - x_t(\beta))/\sigma)}{\Phi(-(y^l - x_t(\beta))/\sigma)}.$$

Ainsi, la fonction de logvraisemblance, qui est la somme sur  $t$  des logarithmes de ces densités est

$$\begin{aligned} \ell(\mathbf{y}, \boldsymbol{\beta}, \sigma) = & -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^n (y_t - x_t(\boldsymbol{\beta}))^2 \\ & - \sum_{t=1}^n \log \left( \Phi \left( -\frac{1}{\sigma} (y^l - x_t(\boldsymbol{\beta})) \right) \right). \end{aligned} \quad (15.43)$$

Les trois premiers termes dans (15.43) composent la fonction de logvraisemblance qui correspond à la régression par moindres carrés non linéaire; consulter l'équation (8.81), par exemple. Cependant, le dernier terme est nouveau. Il s'agit de l'opposé de la somme sur  $t$  des logarithmes des probabilités qu'une observation appartienne à l'échantillon avec la fonction de régression  $x_t(\boldsymbol{\beta})$ . Comme ces probabilités doivent être inférieures à 1, ce terme doit toujours être positif. La présence de ce quatrième terme engendre une divergence entre les estimations de  $\boldsymbol{\beta}$  et  $\sigma$  par ML et leurs analogues par moindres carrés et assure la convergence des estimations ML.

Evidemment ce modèle pourrait facilement être modifié pour permettre d'autres formes de troncature, telles que la troncature des valeurs supérieures ou la troncature à la fois des valeurs inférieures et supérieures. Il peut être révélateur pour les lecteurs d'exécuter la fonction de logvraisemblance pour le modèle de régression (15.41) si l'échantillon est tronqué selon chacune des deux règles suivantes:

$$\begin{aligned} y_t \text{ observé quand } y_t &\leq y^u \text{ et} \\ y_t \text{ observé quand } y^l &\leq y_t \leq y^u, \end{aligned}$$

où  $y^u$  est maintenant une borne supérieure fixée.

Il n'est pas difficile habituellement de maximiser la fonction de logvraisemblance (15.43), en utilisant n'importe laquelle des approches standards. Greene (1990b) aborde certains problèmes qui pourraient potentiellement survenir et montre que, en pratique, la fonction de logvraisemblance aura presque toujours un unique maximum, même si elle n'est pas, en général, globalement concave. La matrice de covariance des estimations ML  $[\boldsymbol{\beta} : \hat{\sigma}]$  sera de dimension  $(k+1) \times (k+1)$ , en supposant que  $\boldsymbol{\beta}$  est un vecteur de dimension  $k$ , et peut comme d'habitude être estimée de diverses manières. Malheureusement, la seule régression artificielle actuellement applicable à ce modèle est la régression OPG. Comme d'habitude, les inférences basées sur cette régression devraient être traitées avec précaution à moins que la taille de l'échantillon ne soit très importante.

Il devrait être clair que la convergence des estimations ML de  $\boldsymbol{\beta}$  et  $\sigma$  obtenues en maximisant (15.43) dépend crucialement des hypothèses que les aléas  $u_t$  dans (15.41) sont réellement normalement, indépendamment,

et identiquement distribués. Autrement, la probabilité que l'élément  $y_t$  soit observé ne sera pas égale à  $\Phi(-(y_t^l - x_t(\beta))/\sigma)$ . Ces hypothèses sont également cruciales pour tous les modèles de régression comprenant des variables dépendantes tronquées ou censurées; voir, par exemple, Hurd (1979) et Arabmazar et Schmidt (1981, 1982). Un certain nombre de techniques a alors été suggéré afin d'obtenir des estimations qui ne sont pas sensibles à ces hypothèses sur la distribution des aléas. Cependant, aucune d'entre elles n'est jusqu'à présent largement utilisée dans les applications économétriques, et entreprendre de discuter de n'importe laquelle de celles-ci nous conduirait bien au-delà des objectifs de cet ouvrage. Consulter, parmi d'autres, Miller (1976), Buckley et James (1979), Powell (1984, 1986), Duncan (1986), Horowitz (1986), Ruud (1986), et Lee (1992).

## 15.7 LES MODÈLES POUR DONNÉES CENSURÉES

Le modèle de régression le plus simple qui comprend une variable dépendante censurée est le **modèle tobit**, ainsi appelé parce qu'il est très étroitement lié au modèle probit et parce qu'il fut proposé à l'origine par Tobin (1958). Une forme simple de modèle tobit est

$$\begin{aligned} y_t^* &= x_t(\beta) + u_t, & u_t &\sim \text{NID}(0, \sigma^2), \\ y_t &= y_t^* \text{ si } y_t^* > 0; & y_t &= 0 \text{ sinon.} \end{aligned} \tag{15.44}$$

Ici  $y_t^*$  est une variable latente observée seulement quand elle est positive. Quand la variable latente est négative, la variable dépendante prend une valeur nulle. La motivation initiale de Tobin était d'étudier les dépenses des ménages sur des biens durables, nulles pour certains ménages et positives pour d'autres.

Il est aisé de modifier le modèle tobit de telle sorte qu'une censure se produise pour d'autres valeurs que zéro, de telle sorte que la censure s'applique à des valeurs supérieures plutôt qu'inférieures, ou de telle sorte que la valeur sur laquelle se produit la censure change (d'une manière non stochastique) sur l'échantillon. Par exemple,  $y_t^*$  pourrait être la demande de places sur un vol d'une compagnie aérienne,  $y_t^c$  pourrait être la capacité de l'appareil (qui pourrait varier sur l'échantillon selon les différents appareils utilisés lors des vols), et  $y_t$  pourrait être le nombre de sièges véritablement occupés. Ainsi, la seconde ligne de (15.44) serait remplacée par

$$y_t = y_t^* \text{ si } y_t^* < y_t^c; \quad y_t = y_t^c \text{ sinon.}$$

Le modèle tobit a été très largement usité dans les travaux appliqués. Des applications de celui-ci ont concerné des sujets très divers tels que le chômage (Ashenfelter et Ham, 1979), l'âge espéré de la retraite (Kotlikoff, 1979), la demande de cuivre (MacKinnon et Olewiler, 1980), et même le nombre d'aventures extra-conjugales (Fair, 1978).



**Tableau 15.1** Non Convergence Causée par la Troncature et la Censure

$y'$	fraction $< y'$	$\text{plim}(\hat{\beta}_0)$	$\text{plim}(\hat{\beta}_1)$	$\text{plim}(\tilde{\beta}_0)$	$\text{plim}(\tilde{\beta}_1)$
0.0	0.076	1.26	0.77	1.07	0.93
0.5	0.167	1.48	0.63	1.18	0.83
1.0	0.316	1.77	0.49	1.37	0.69
1.5	0.500	2.12	0.37	1.67	0.50
2.0	0.684	2.51	0.28	2.06	0.31
2.5	0.833	2.93	0.21	2.51	0.16

Il est tout aussi incorrect d'utiliser une régression par moindres carrés avec des données censurées qu'avec des données tronquées. Le Tableau 15.1 contient certains résultats numériques pour l'estimation OLS du modèle

$$y_t^* = \beta_0 + \beta_1 x_t + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

où  $y_t$  provient de  $y_t^*$  soit par troncature soit par censure sur des valeurs inférieures à  $y'$ . Pour cette illustration, les véritables valeurs de  $\beta_0$ ,  $\beta_1$ , et  $\sigma$  étaient toutes unitaires, et  $x_t$  était uniformément distribuée sur l'intervalle (0,1). Chaque ligne du tableau correspond à une valeur différente de  $y'$  et en conséquence à une proportion différente des observations limitées. Les estimations basées sur l'échantillon tronqué sont désignées par  $\hat{\beta}_0$  et  $\hat{\beta}_1$ , tandis que celles basées sur l'échantillon censuré sont désignées par  $\tilde{\beta}_0$  et  $\tilde{\beta}_1$ .<sup>8</sup>

Il apparaît à partir des résultats du tableau que la non convergence due à la troncature ou à la censure peut être très importante, la troncature provoquant une non convergence encore plus prononcée que la censure, du moins dans cet exemple. Comme nous pouvions nous y attendre, la non convergence augmente avec la proportion des observations limitées. Notons que pour le cas censuré,  $\text{plim}(\tilde{\beta}_1)/\beta_1$  est essentiellement égale à la proportion des observations non limitées dans l'échantillon,  $1 - \Pr(y_t < y')$ . Greene (1981a) a dérivé ce résultat de façon analytique pour tous les coefficients de pente dans un modèle de régression linéaire, sous l'hypothèse particulière que les régresseurs sont normalement distribués. Ceci semble fournir une très bonne approximation pour certains autres cas, dont celui analysé dans le tableau.

<sup>8</sup> Ces résultats ont été obtenus au moyen d'une expérience Monte Carlo qui comprenait 500 simulations, avec chacune 50,000 observations. Bien que l'erreur expérimentale devrait être très petite, les derniers chiffres reportés dans le tableau peuvent ne pas être exacts. Par exemple, il est aisé de voir que dans cet exemple la fraction tronquée ou censurée quand  $y'$  est 1.5 doit être 0.50, et c'est le nombre reporté dans le tableau. Cependant, le nombre effectivement observé dans les expériences a été 0.498.

Le modèle tobit est habituellement estimé par maximum de vraisemblance. Par souci de simplicité, nous discuterons de l'estimation du modèle tobit simple donné par (15.44). Nous voyons immédiatement que

$$\begin{aligned}\Pr(y_t = 0) &= \Pr(y_t^* \leq 0) = \Pr(x_t(\beta) + u_t \leq 0) \\ &= \Pr\left(\frac{u_t}{\sigma} \leq -\frac{x_t(\beta)}{\sigma}\right) = \Phi\left(-\frac{1}{\sigma}x_t(\beta)\right).\end{aligned}$$

Ainsi, la contribution des observations du type  $y_t = 0$  à la fonction de logvraisemblance est

$$\ell_t(y_t, \beta, \sigma) = \log\left(\Phi\left(-\frac{1}{\sigma}x_t(\beta)\right)\right). \quad (15.45)$$

Lorsque  $y_t$  est positive, sa densité est

$$\frac{\sigma^{-1}\phi((y_t - x_t(\beta))/\sigma)}{\Pr(y_t > 0)}. \quad (15.46)$$

Cependant, la contribution des observations du type  $y_t > 0$  à la fonction de logvraisemblance n'est pas le logarithme de (15.46), parce que ces observations ne surviennent seulement qu'avec une probabilité  $\Pr(y_t > 0)$ . En multipliant (15.46) par  $\Pr(y_t > 0)$  et en calculant le logarithme nous obtenons

$$\log\left(\frac{1}{\sigma}\phi\left(\frac{1}{\sigma}(y_t - x_t(\beta))\right)\right), \quad (15.47)$$

qui est la contribution d'une observation à la fonction de logvraisemblance dans un modèle de régression sans censure.

La fonction de logvraisemblance pour le modèle tobit est ainsi

$$\sum_{y_t=0} \log\left(\Phi\left(-\frac{1}{\sigma}x_t(\beta)\right)\right) + \sum_{y_t>0} \log\left(\frac{1}{\sigma}\phi\left(\frac{1}{\sigma}(y_t - x_t(\beta))\right)\right). \quad (15.48)$$

Le premier terme est simplement la somme sur les observations limitées de l'expression (15.45), et le second est la somme sur les observations non limitées de l'expression (15.47). Le premier terme semble correspondre à la fonction de logvraisemblance pour un modèle probit. Nous pouvons montrer cette similitude en linéarisant la fonction de régression et en imposant la normalisation  $\sigma = 1$ , auquel cas  $\Phi(-x_t(\beta)/\sigma)$  devient  $1 - \Phi(\mathbf{X}_t\beta)$ , et en comparant la fonction de logvraisemblance à (15.09). Par contraste, le second terme dans (15.48) ressemble à la fonction de logvraisemblance pour un modèle de régression non linéaire.

Les lecteurs peuvent s'interroger à raison sur la validité de cette fonction de logvraisemblance. Après tout, le premier terme est une somme de logarithmes d'un certain nombre de probabilités, tandis que le second terme est

une somme de logarithmes d'un certain nombre de densités. Ce mélange relativement étrange provient du fait que la variable dépendante dans un modèle tobit est parfois une variable aléatoire discrète (pour les observations limitées) et parfois une variable continue (pour les observations non limitées). En raison de ce mélange de variables aléatoires discrètes et continues, les démonstrations standards de la convergence et de la normalité asymptotique des estimateurs ML ne s'appliquent pas au modèle tobit. Cependant, Amemiya (1973c), dans un article bien connu, a montré que l'estimateur ML possède en réalité toutes les propriétés asymptotiques habituelles. Il fournit également les expressions pour les éléments de la matrice d'information.

Il n'est pas difficile de maximiser la fonction de logvraisemblance (15.48). Bien qu'elle ne soit pas globalement concave dans sa paramétrisation naturelle, Olsen (1978) a montré que quand  $x_t(\beta) = \mathbf{X}_t\beta$ , celle-ci ne possède pas un maximum unique. L'argument clé est que nous pouvons reparamétriser le modèle en termes des paramètres  $\alpha \equiv \beta/\sigma$  et  $h \equiv 1/\sigma$ , et montrer que la fonction de logvraisemblance peut être globalement concave avec cette dernière paramétrisation. Ceci implique que celle-ci doit avoir un unique maximum quelle que soit la paramétrisation. La matrice de covariance de dimension  $(k+1) \times (k+1)$  des estimations ML peut être estimée habituellement de plusieurs manières. Malheureusement, comme avec le modèle de régression tronqué examiné dans la section précédente, la seule régression artificielle applicable à ce modèle est la régression OPG.

Il existe une relation intéressante entre les modèles tobit, de régression tronquée et probit. Supposons, pour faire simple, que  $x_t(\beta) = \mathbf{X}_t\beta$ . Alors la fonction de logvraisemblance du modèle tobit peut être réécrite comme

$$\sum_{y_t > 0} \log\left(\frac{1}{\sigma} \phi\left(\frac{1}{\sigma}(y_t - \mathbf{X}_t\beta)\right)\right) + \sum_{y_t = 0} \log\left(\Phi\left(-\frac{1}{\sigma} \mathbf{X}_t\beta\right)\right). \quad (15.49)$$

Additionnons et soustrayons maintenant le terme  $\sum_{y_t > 0} \log(\Phi(\mathbf{X}_t\beta/\sigma))$  à (15.49), qui devient alors

$$\begin{aligned} & \sum_{y_t > 0} \log\left(\frac{1}{\sigma} \phi\left(\frac{1}{\sigma}(y_t - \mathbf{X}_t\beta)\right)\right) - \sum_{y_t > 0} \log\left(\Phi\left(\frac{1}{\sigma} \mathbf{X}_t\beta\right)\right) \\ & + \sum_{y_t = 0} \log\left(\Phi\left(-\frac{1}{\sigma} \mathbf{X}_t\beta\right)\right) + \sum_{y_t > 0} \log\left(\Phi\left(\frac{1}{\sigma} \mathbf{X}_t\beta\right)\right). \end{aligned} \quad (15.50)$$

Ici, la première ligne est la fonction de logvraisemblance pour un modèle de régression tronqué; il s'agit simplement de (15.43) avec  $y^l = 0$  et  $x_t(\beta) = \mathbf{X}_t\beta$  et avec un ensemble d'observations auxquelles s'appliquent les sommations ajusté convenablement. La seconde ligne est la fonction de logvraisemblance pour un modèle probit avec la fonction indice  $\mathbf{X}_t\beta/\sigma$ . Naturellement, si seule la seconde ligne apparaissait, nous ne pourrions pas identifier  $\beta$  et  $\sigma$

séparément, mais comme nous disposons également de la première ligne, ceci ne constitue pas un problème.

L'expression (15.50) montre clairement que le modèle tobit est comparable à un modèle de régression tronqué combiné à un modèle probit, les vecteurs de coefficients des deux derniers modèles étant contraints à être proportionnels. Cragg (1971) a avancé l'idée que cette restriction peut parfois être irréaliste et a proposé plusieurs modèles plus généraux comme alternatives plausibles au modèle tobit. Il peut parfois être souhaitable de tester le modèle tobit contre un ou plus de ces modèles généraux; voir Lin et Schmidt (1984) et Greene (1990a, Chapitre 21).

Comme nous l'avons mentionné plus tôt, il est facile de modifier le modèle tobit pour gérer différents types de censures. Par exemple, une possibilité est un modèle avec une **double censure**. Supposons que

$$y_t^* = x_t(\beta) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

$$y_t = y_t^* \text{ si } y_t^l \leq y_t^* \leq y_t^u; \quad y_t = y_t^l \text{ si } y_t^* < y_t^l; \quad y_t = y_t^u \text{ si } y_t^* > y_t^u.$$

Ce modèle a été étudié par Rosett et Nelson (1975) et Nakamura et Nakamura (1983), parmi d'autres. Nous voyons que la fonction de logvraisemblance est

$$\begin{aligned} & \sum_{y_t^l \leq y_t^* \leq y_t^u} \log \left( \frac{1}{\sigma} \phi \left( \frac{1}{\sigma} (y_t - \mathbf{X}_t \beta) \right) \right) + \sum_{y_t^* < y_t^l} \log \left( \Phi \left( \frac{1}{\sigma} (y_t^l - \mathbf{X}_t \beta) \right) \right) \\ & + \sum_{y_t^* > y_t^u} \log \left( \Phi \left( -\frac{1}{\sigma} (y_t^u - \mathbf{X}_t \beta) \right) \right). \end{aligned} \quad (15.51)$$

Le premier terme correspond aux observations non limitées, le deuxième aux observations fixées à la limite inférieure  $y_t^l$ , et le troisième aux observations fixées à la limite supérieure  $y_t^u$ . La maximisation de (15.51) est immédiate.

De nombreux autres modèles pour régression avec données censurées et tronquées furent proposés dans la littérature. Certains d'entre eux traitent de situations dans lesquelles il existe deux ou plusieurs variables dépendantes jointes. Nelson et Olsen (1978) et Lee (1981) sont des références importantes; consulter les synthèses de Amemiya (1985, Chapitre 10) et Dhrymes (1986). Nous n'avons pas suffisamment de place pour discuter de cette littérature plus en détails. Cependant, il est utile de mentionner un cas fréquent.

Supposons que  $y_t^*$  soit une variable latente déterminée par le modèle

$$y_t^* = \mathbf{X}_t \beta + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad (15.52)$$

et que  $y_t$  provienne de  $y_t^*$  par une forme quelconque de censure ou de troncature. Il s'ensuit que le modèle qui est réellement estimé est un modèle probit, tobit, ou de régression tronquée. C'est évidemment le type de troncature ou de censure pour passer de  $y_t^*$  vers  $y_t$  qui déterminera lequel de ces

modèles est approprié. Supposons maintenant que une (plusieurs) variable(s) dépendante(s) dans le vecteur  $\mathbf{X}_t$  puisse(nt) être corrélée(s) avec les aléas  $u_t$ . Si tel est le cas, les estimations ML habituelles de  $\beta$  seront évidemment non convergentes.

Heureusement, il est très facile de tester une non convergence liée à la possible corrélation entre certaines variables indépendantes et les aléas dans (15.52). Le test est très similaire au test DWH pour la non convergence causée par une possible endogénéité, test discuté dans la Section 7.9. Supposons que  $\mathbf{W}$  soit une matrice de variables instrumentales comprenant toutes les colonnes de  $\mathbf{X}$  (une matrice avec comme ligne type  $\mathbf{X}_t$ ) dont nous savons qu'elles sont exogènes ou prédéterminées. Pour exécuter le test, nous régressons tout d'abord les colonnes restantes de  $\mathbf{X}$ , disons  $\mathbf{X}^*$ , sur  $\mathbf{W}$  et sauvegardons les résidus  $\mathbf{M}_W \mathbf{X}^*$ . Nous calculons alors soit un test LR soit un test LM pour l'hypothèse que  $\gamma = \mathbf{0}$  dans le modèle à variable latente fictif

$$y_t^* = \mathbf{X}_t \beta + (\mathbf{M}_W \mathbf{X}^*)_t \gamma + u_t, \quad u_t \sim \text{NID}(0, \sigma^2).$$

Ici  $(\mathbf{M}_W \mathbf{X}^*)_t$  sert d'estimation des parties stochastiques des variables probablement endogènes dans  $\mathbf{X}_t$ . Si ces variables ne sont pas corrélées avec  $u_t$ , et que le modèle à variable latente est spécifié correctement, le vecteur  $\gamma$  devrait être nul. Ce test fut utilisé par MacKinnon et Olewiler (1980) et détaillé ensuite par Smith et Blundell (1986).

Il est entendu qu'à la fois les modèles correspondant aux hypothèses nulle et alternative pour ce test sont véritablement des modèles de régression tronquée, probit, ou tobit, qui dépendent de la manière dont  $y_t$  est obtenue à partir de  $y_t^*$ . Comme d'habitude, les tests LM peuvent être basés sur les régressions artificielles. Comme seule la régression OPG est disponible pour les modèles tobit et de régression tronqués, il peut être préférable d'utiliser un test LR dans ces cas. Quand nous estimons le modèle alternatif, il s'avère que les estimations de  $\beta$  sont convergentes même si l'hypothèse nulle est fausse, comme elles l'étaient dans le cas de la régression linéaire examiné dans la Section 7.9. Cependant, la matrice de covariance ordinaire produite par cette procédure n'est pas valable asymptotiquement quand  $\gamma \neq \mathbf{0}$ , pour la même raison qu'elle ne l'était pas dans le cas de la régression linéaire.

## 15.8 SÉLECTION D'ÉCHANTILLON

Dans la Section 15.6, nous avons discuté des modèles dans lesquels l'échantillon avait été tronqué selon la valeur de la variable dépendante. Cependant, dans de nombreux cas pratiques, la troncature n'est pas basée sur la valeur de la variable dépendante mais plutôt sur la valeur d'une autre variable qui lui est corrélée avec elle. Par exemple, les gens peuvent choisir d'entrer sur le marché du travail seulement si leur salaire de marché excède leur salaire de réserve. Ainsi l'échantillon des gens qui sont sur le marché du travail

exclura ceux pour qui le salaire de réserve excède leur salaire de marché. Si la variable dépendante est un élément corrélé avec leurs salaires de réserve ou de marché, l'utilisation des moindres carrés fournira des estimations non convergentes. Dans ce cas, l'échantillon est dit **sélectionné** sur la base de la différence entre le salaire de réserve et le salaire de marché, et le problème que ce type de sélection provoque est souvent désigné sous le nom de **biais de sélection d'échantillon**. Heckman (1974, 1976, 1979), Hausman et Wise (1977), et Lee (1978) sont les tous premiers articles sur ce sujet.

La meilleure manière de comprendre les caractéristiques clés des modèles impliquant la sélection d'échantillon consiste à examiner un modèle simple en détail. Supposons que  $y_t^*$  et  $z_t^*$  soient deux variables latentes, générées par le processus bivarié

$$\begin{bmatrix} y_t^* \\ z_t^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_t \boldsymbol{\beta} \\ \mathbf{W}_t \boldsymbol{\gamma} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}, \quad \begin{bmatrix} u_t \\ v_t \end{bmatrix} \sim \text{NID} \left( \mathbf{0}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right), \quad (15.53)$$

où  $\mathbf{X}_t$  et  $\mathbf{W}_t$  sont des vecteurs d'observations sur les variables exogènes ou prédéterminées,  $\boldsymbol{\beta}$  et  $\boldsymbol{\gamma}$  sont des vecteurs paramétriques inconnus,  $\sigma$  est l'écart type de  $u_t$  et  $\rho$  représente la corrélation entre  $u_t$  et  $v_t$ . La restriction que la variance de  $v_t$  est égale à 1 est imposée parce que seul le signe de  $z_t^*$  sera observé. De fait, les variables réellement observées sont  $y_t$  et  $z_t$ , et elles sont reliées à  $y_t^*$  et  $z_t^*$  comme suit:

$$\begin{aligned} y_t &= y_t^* \text{ si } z_t^* > 0; \quad y_t = 0 \text{ sinon;} \\ z_t &= 1 \text{ si } z_t^* > 0; \quad z_t = 0 \text{ sinon.} \end{aligned}$$

Il existe deux types d'observations: celles pour lesquelles à la fois  $y_t$  et  $z_t$  sont nulles et celles pour lesquelles  $z_t = 1$  et  $y_t$  est égale à  $y_t^*$ . La fonction de logvraisemblance pour ce modèle est ainsi

$$\sum_{z_t=0} \log(\Pr(z_t = 0)) + \sum_{z_t=1} \log(\Pr(z_t = 1) f(y_t^* | z_t = 1)), \quad (15.54)$$

où  $f(y_t^* | z_t = 1)$  désigne la densité de  $y_t^*$  conditionnelle à  $z_t = 1$ . Le premier terme de (15.54) est la somme sur les observations pour lesquelles  $z_t = 0$  des logarithmes de la probabilité que  $z_t = 0$ . C'est exactement le même terme que celui qui correspond à  $z_t$  par lui-même dans un modèle probit. Le second terme est la somme sur les observations pour lesquelles  $z_t = 1$  de la probabilité que  $z_t = 1$  fois la densité de  $y_t$  conditionnelle à  $z_t = 1$ . En utilisant le fait que nous pouvons factoriser une densité jointe de n'importe quelle manière, le second terme peut aussi être écrit comme

$$\sum_{z_t=1} \log(\Pr(z_t = 1 | y_t^*) f(y_t^*)),$$

où  $f(y_t^*)$  est la densité conditionnelle de  $y_t^*$ , qui est simplement une densité normale d'espérance conditionnelle  $\mathbf{X}_t\boldsymbol{\beta}$  et de variance  $\sigma^2$ .

La seule difficulté dans l'écriture explicite de la fonction de logvraisemblance (15.54) est de calculer  $\Pr(z_t = 1 | y_t^*)$ . Comme  $u_t$  et  $v_t$  sont normaux bivariés, nous pouvons écrire

$$z_t^* = \mathbf{W}_t\boldsymbol{\gamma} + \rho\left(\frac{1}{\sigma}(y_t^* - \mathbf{X}_t\boldsymbol{\beta})\right) + \varepsilon_t, \quad \varepsilon_t \sim \text{NID}(0, (1 - \rho^2)).$$

Il s'ensuit que

$$\Pr(z_t = 1) = \Phi\left(\frac{\mathbf{W}_t\boldsymbol{\gamma} + \rho((y_t - \mathbf{X}_t\boldsymbol{\beta})/\sigma)}{(1 - \rho^2)^{1/2}}\right),$$

comme  $y_t = y_t^*$  quand  $z_t = 1$ . Ainsi la fonction de logvraisemblance (15.54) devient

$$\begin{aligned} & \sum_{z_t=0} \log(\Phi(-\mathbf{W}_t\boldsymbol{\gamma})) + \sum_{z_t=1} \log\left(\frac{1}{\sigma}\phi(y_t - \mathbf{X}_t\boldsymbol{\beta})\right) \\ & + \sum_{z_t=1} \log\left(\Phi\left(\frac{\mathbf{W}_t\boldsymbol{\gamma} + \rho((y_t - \mathbf{X}_t\boldsymbol{\beta})/\sigma)}{(1 - \rho^2)^{1/2}}\right)\right). \end{aligned} \quad (15.55)$$

Le premier terme ressemble au terme correspondant pour un modèle probit. Le deuxième terme ressemble à la fonction de logvraisemblance pour un modèle de régression linéaire à erreurs normales. Le troisième terme est celui que nous n'avons pas vu auparavant.

Les estimations par maximum de vraisemblance peuvent être obtenues de manière habituelle par la maximisation de (15.55). Cependant, cette maximisation est relativement onéreuse, et une technique de calcul plus simple proposée par Heckman (1976) est souvent utilisée à la place d'une estimation ML. La **méthode en deux étapes de Heckman** est basée sur le fait que la première équation de (15.53) peut être réécrite comme

$$y_t^* = \mathbf{X}_t\boldsymbol{\beta} + \rho\sigma v_t + e_t. \quad (15.56)$$

L'idée est de remplacer  $y_t^*$  par  $y_t$  et  $v_t$  par son espérance conditionnelle à  $z_t = 1$  et à la valeur réalisée de  $\mathbf{W}_t\boldsymbol{\gamma}$ . Comme nous l'avons vu à partir de (15.42), cette espérance conditionnelle est  $\phi(\mathbf{W}_t\boldsymbol{\gamma})/\Phi(\mathbf{W}_t\boldsymbol{\gamma})$ , une quantité parfois désignée sous le nom de **ratio inverse de Mills**. En conséquence, la régression (15.56) devient

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \rho\sigma \frac{\phi(\mathbf{W}_t\boldsymbol{\gamma})}{\Phi(\mathbf{W}_t\boldsymbol{\gamma})} + \text{résidu}. \quad (15.57)$$

Il est maintenant facile de voir comment opère la méthode en deux étapes de Heckman. Dans une première étape, un modèle probit ordinaire est utilisé

pour obtenir des estimations convergentes  $\hat{\gamma}$  des paramètres de l'équation de sélection. Ensuite, dans une seconde étape, le **régresseur de sélection**  $\phi(\mathbf{W}_t\gamma)/\Phi(\mathbf{W}_t\gamma)$  est évalué en  $\hat{\gamma}$  et la régression (15.57) est estimée par OLS à l'aide des observations pour lesquelles  $y_t > 0$ . Cette régression fournit aussi bien un test pour la sélection d'échantillon qu'une technique d'estimation. Le coefficient du régresseur de sélection est  $\rho\sigma$ . Comme  $\sigma \neq 0$ , le  $t$  de Student ordinaire pour la nullité de ce coefficient peut être utilisé pour tester l'hypothèse que  $\rho = 0$ ; celle-ci sera asymptotiquement distribuée selon la  $N(0, 1)$  sous l'hypothèse nulle. Ainsi, si ce coefficient n'est pas significativement différent de zéro, l'expérimentateur peut raisonnablement décider que la sélection n'est pas un problème pour cet ensemble de données, et continuer à utiliser les moindres carrés comme d'habitude.

Même quand l'hypothèse que  $\rho = 0$  ne peut pas être acceptée, l'estimation OLS de la régression (15.57) fournit des estimations convergentes de  $\beta$ . Cependant, la matrice de covariance OLS n'est valable que lorsque  $\rho = 0$ . A cet égard, la situation est très similaire à celle rencontrée à la fin de la section précédente, quand nous testions des biais de simultanéité potentiels dans des modèles à variables dépendantes tronquées ou censurées. Il existe en réalité deux problèmes. Tout d'abord, les résidus dans (15.57) seront hétéroscédastiques, puisqu'un résidu type est égal à

$$u_t - \rho\sigma \frac{\phi(\mathbf{W}_t\gamma)}{\Phi(\mathbf{W}_t\gamma)}.$$

Ensuite, le régresseur de sélection est traité comme n'importe quel autre régresseur, quand il s'agit en réalité d'une partie de l'aléa. Nous pourrions résoudre le premier problème en utilisant un estimateur de matrice de covariance robuste à l'hétéroscédasticité (voir le Chapitre 16), mais cela ne résoudra pas le second problème. Il est possible d'obtenir une estimation valable de la matrice de covariance compatible avec les estimations en deux étapes de  $\beta$  à partir de (15.57). Cependant, le calcul est peu pratique, et la matrice de covariance estimée n'est pas toujours définie positive. Consulter Greene (1981b) et Lee (1982) pour plus de détails.

Il faut insister sur le fait que la convergence de cet estimateur en deux étapes, comme celle de l'estimateur ML, dépend de façon critique de l'hypothèse de normalité. Nous pouvons comprendre cela à partir de la spécification du régresseur de sélection comme l'inverse du ratio Mills  $\phi(\mathbf{W}_t\gamma)/\Phi(\mathbf{W}_t\gamma)$ . Quand les éléments de  $\mathbf{W}_t$  sont identiques aux éléments de  $\mathbf{X}_t$ , ou en sont un sous-ensemble, comme c'est souvent le cas dans la pratique, c'est seulement la non linéarité de  $\phi(\mathbf{W}_t\gamma)/\Phi(\mathbf{W}_t\gamma)$  comme fonction de  $\mathbf{W}_t\gamma$  qui identifie les paramètres de la seconde étape. La forme exacte de la relation non linéaire dépend de façon critique de l'hypothèse de normalité. Pagan et Vella (1989), Smith (1989), et Peters et Smith (1991) discutent de diverses manières de tester cette hypothèse cruciale. Beaucoup des tests suggérés par ces auteurs sont des applications de la régression OPG.



Bien que la méthode en deux étape pour traiter la sélection d'échantillon soit largement utilisée, notre recommandation serait d'utiliser la régression (15.57) seulement comme procédure pour tester l'hypothèse nulle d'absence de biais de sélection n'est pas présent. Quand cette hypothèse nulle est rejetée, nous préfererons probablement utiliser l'estimation ML basée sur (15.55) plutôt que la méthode en deux étape, à moins que son calcul ne soit prohibitif.

## 15.9 CONCLUSION

Notre traitement des modèles à réponse binaire dans les Sections 15.2 à 15.4 a été raisonnablement détaillé, mais les discussions plus générales des modèles à réponse qualitative et des modèles à variable dépendante limitée ont été nécessairement très superficielles. Le praticien qui a l'intention de réaliser un travail empirique qui emploie ce type de modèle souhaitera consulter certaines synthèses plus fournies dont nous avons donné les références. Toutes les méthodes pour traiter des variables dépendantes limitées dont nous avons discuté reposent lourdement sur les hypothèses de normalité et d'homoscédasticité. Il faudrait toujours tester ces hypothèses. Un certain nombre de méthodes pour réaliser ces tests de la sorte a été proposées; consulter, parmi d'autres, Bera, Jarque, et Lee (1984), Lee et Maddala (1985), Blundell (1987), Chesher et Irish (1987), Pagan et Vella (1989), Smith (1989), et Peters et Smith (1991).

## TERMES ET CONCEPTS

biais de sélection d'échantillon	modèle logit
classificatrices parfaites	modèle logit conditionnel
données censurées	modèle logit multinomial (ou
données tronquées	multiple)
double censure	modèle logit ordonné
échantillon sélectionné	modèle probit
fonction de transformation	modèle probit ordonné
fonction indice	modèle tobit
fonction logistique	modèles à variable dépendante limitée
indépendance par rapport aux	modèles à variable latente
alternatives non pertinentes (IIA)	ratio inverse de Mills
méthode en deux étapes de Heckman	régression de modèle à réponse binaire
modèles à réponse binaire (ou choix	(BRMR)
binaire)	régresseur de sélection
modèle à réponse qualitative	réponses non ordonnées et ordonnées
modèle à réponse qualitative ordonné	variable dépendante binaire
modèle de probabilité linéaire	variables dépendantes limitées
modèle de régression tronquée	variables latentes

# Chapitre 16

## L'Hétéroscédasticité

### 16.1 INTRODUCTION

La plupart des résultats obtenus jusqu'à présent pour les modèles de régression reposaient explicitement ou implicitement sur l'hypothèse d'homoscédasticité des aléas, et certains résultats dépendaient sur l'hypothèse supplémentaire qu'ils sont normalement distribués. Cependant, dans la pratique, les deux hypothèses d'homoscédasticité et de normalité semblent souvent être enfreintes. C'est principalement le cas lorsque les données sont relatives à des observations en coupe transversale des ménages ou des entreprises ou à des observations chronologiques des actifs financiers. Dans ce chapitre, nous traitons d'un certain nombre de thèmes importants, connexes à l'hétéroscédasticité, à la non normalité et à d'autres défaillances des hypothèses habituelles concernant les aléas des modèles de régression.

Comme nous l'avons vu dans le Chapitre 9, il est parfaitement facile d'estimer un modèle de régression par moindres carrés pondérés (c'est-à-dire par GLS) quand les aléas sont hétéroscédastiques avec une structure d'hétéroscédasticité déterminée par une fonction scédastique connue. Nous avons aussi vu qu'il était raisonnablement facile d'estimer un modèle de régression par GLS faisables ou par maximum de vraisemblance lorsque la forme de la fonction scédastique est connue mais pas ses paramètres. De plus, comme nous l'avons vu dans le Chapitre 14, faire subir à la variable dépendante (et probablement également à la fonction de régression) une transformation non linéaire appropriée peut éliminer entièrement l'hétéroscédasticité. Bien que parfois très efficaces et pratiques, ces techniques ne nous permettent pas de gérer le cas le plus commun où presque rien n'est connu sur la fonction scédastique.

Dans la Section 16.2, nous discutons des propriétés des estimations NLS (et OLS) lorsque les aléas sont hétéroscédastiques. Sous des hypothèses raisonnables, les estimations demeurent convergentes et asymptotiquement normales, mais leur matrice de covariance asymptotique diffère de la matrice habituelle. Dans la Section 16.3, nous montrons alors qu'il est possible d'employer un **estimateur de la matrice de covariance robuste à l'hétéroscédasticité**, même si pratiquement rien n'est connu sur la forme

de la fonction scédastique. Ce résultat extrêmement important nous permet de réaliser des inférences asymptotiquement valables sur des modèles de régression linéaire et non linéaire sous des conditions assez peu contraignantes. Cela fournit également une justification à la régression de Gauss-Newton robuste à l'hétéroscédasticité discutée dans la Section 11.6.

Dans la Section 16.4, nous discutons de l'idée d'**hétéroscédasticité conditionnelle autorégressive**, ou **ARCH**, qui s'est avérée extrêmement utile dans la modélisation des aléas associés aux modèles de régression pour certains types de données temporelles, et en particulier pour les données des marchés financiers. Puis, dans les Sections 16.5 et 16.6, nous discutons de certains aspects des tests d'hétéroscédasticité qui échappaient à la discussion de la Section 11.5. En particulier, nous discutons des conséquences de l'orthogonalité des directions de régression et des directions scédastiques, quel que soit le modèle dont les aléas sont distribués indépendamment de la fonction de régression.

Dans la Section 16.7, nous portons notre attention sur les tests de normalité des aléas, et focalisons sur les tests d'asymétrie et d'excès de kurtosis. Il s'avère que les tests de normalité sont très faciles dans le contexte des modèles de régression. Dans la section suivante, nous introduisons une classe assez large de tests appelés **tests de moments conditionnels**. Ces tests sont étroitement liés aux **tests de la matrice d'information**, discutés dans la Section 16.9.

## 16.2 MOINDRES CARRÉS ET HÉTÉROSCÉDASTICITÉ

Les propriétés des estimateurs des moindres carrés ordinaires et non linéaires appliqués aux modèles à erreurs hétéroscédastiques sont très similaires. Pour simplifier, nous commençons donc par le cas linéaire. Supposons que le modèle de régression linéaire estimé soit

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

où  $\mathbf{X}$  est une matrice de dimension  $n \times k$  répondant à la condition de régularité asymptotique usuelle que  $n^{-1}\mathbf{X}^\top \mathbf{X}$  a pour limite une matrice définie positive  $O(1)$ . Les données sont en réalité générées par

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \boldsymbol{\Omega}, \quad (16.01)$$

où  $\boldsymbol{\Omega}$  est une matrice diagonale avec des éléments types diagonaux  $\omega_t^2$  bornés supérieurement et inférieurement. Nous nous intéressons aux propriétés de l'estimateur OLS  $\hat{\boldsymbol{\beta}}$  lorsque le DGP est (16.01). Clairement,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}. \quad (16.02)$$

Il s'ensuit que

$$\text{plim}_{n \rightarrow \infty}(\hat{\beta}) = \beta_0 + \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{u} \right).$$

Il est par conséquent clair que  $\hat{\beta}$  sera une estimation convergente de  $\beta$  à condition que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{u} \right) = \mathbf{0}.$$

Comme nous l'avons vu dans la Section 9.5, cette condition ne tient pas toujours quand les aléas ne sont pas i.i.d. Bien que la discussion n'ait pas été rigoureuse, il était clair que trois situations devaient être éliminées. Deux de celles-ci impliquaient des matrices  $\Omega$  non diagonales, la troisième impliquait des variances non bornées. Puisque ces trois situations sont exclues par les hypothèses déjà formulées, nous pouvons sans doute conclure que  $\hat{\beta}$  est effectivement convergent. Pour un traitement beaucoup plus complet de ce sujet, voir White (1984).

Si  $\mathbf{X}$  peut être considérée comme fixe, il est facile de voir à partir de (16.02) que

$$\begin{aligned} \mathbf{V}(\hat{\beta} - \beta_0) &= E \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \right) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \Omega \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned} \quad (16.03)$$

Nous ferons référence à la dernière expression en tant que **matrice de covariance des OLS généralisés**. Nous pouvons la comparer à la matrice de covariance habituelle des OLS,

$$\sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1}, \quad (16.04)$$

où  $\sigma_0^2$  serait dans ce cas la limite en probabilité de la moyenne des  $\omega_t^2$ , et avec la matrice de covariance GLS

$$(\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1}.$$

Le Théorème de Gauss-Markov (Théorème 5.3) implique que

$$(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \Omega \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1}$$

doit être une matrice semi-définie positive. Elle sera une matrice nulle dans les circonstances (relativement rares) où le Théorème de Kruskal s'applique et les estimations OLS et GLS coïncident (voir la Section 9.3).

Evidemment, deux problèmes différents surgissent si nous utilisons les OLS en lieu et place des moindres carrés pondérés, ou GLS. Le premier est que les estimations OLS ne seront pas efficaces, ce qui découle du Théorème de

**Tableau 16.1** Ecarts Types Corrects et Incorrects

$\alpha$	$\hat{\beta}_0$ (Incorrect)	$\tilde{\beta}_0$ (Correct)	$\tilde{\beta}_0$	$\hat{\beta}_1$ (Incorrect)	$\tilde{\beta}_1$ (Correct)	$\tilde{\beta}_1$
0.5	0.164	0.134	0.110	0.285	0.277	0.243
1.0	0.142	0.101	0.048	0.246	0.247	0.173
1.5	0.127	0.084	0.019	0.220	0.231	0.136
2.0	0.116	0.074	0.0073	0.200	0.220	0.109
2.5	0.107	0.068	0.0030	0.185	0.212	0.083
3.0	0.100	0.064	0.0013	0.173	0.206	0.056
3.5	0.094	0.061	0.0007	0.163	0.200	0.033
4.0	0.089	0.059	0.0003	0.154	0.195	0.017

Gauss-Markov. Le second est que la matrice de covariance standard des OLS ne sera pas équivalente, dans la plupart des cas, à la matrice de covariance des OLS généralisés, l'expression la plus à droite dans (16.03). La gravité de chacun de ces problèmes dépendra évidemment des formes exactes de  $\mathbf{X}$  et  $\mathbf{\Omega}$ .

Il peut être intéressant d'examiner un exemple numérique. Le modèle est

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

et le DGP est

$$y_t = 1 + x_t + u_t, \quad u_t \sim N(0, x_t^\alpha),$$

avec  $n = 100$ ,  $x_t$  est uniformément distribuée entre 0 et 1, et  $\alpha$  est un paramètre qui prend différentes valeurs. Le Tableau 16.1 rassemble les écarts types des estimations OLS  $\hat{\beta}_0$  et  $\hat{\beta}_1$ , et des estimations GLS  $\tilde{\beta}_0$  et  $\tilde{\beta}_1$ . Pour les estimations OLS, le tableau illustre à la fois les écarts types corrects et incorrects provenant de la formule habituelle.<sup>1</sup>

Malgré leur simplicité, les résultats du Tableau 16.1 illustrent deux faits. Tout d'abord, les GLS peuvent n'être que légèrement plus efficaces que les OLS lorsque  $\alpha = 0.5$ , ou beaucoup plus efficaces pour des valeurs plus grandes que  $\alpha$ . Par ailleurs, les écarts types des OLS habituels peuvent être soit trop grands (comme ils le sont toujours pour  $\beta_0$ ) soit trop petits (comme ils le sont toujours pour  $\beta_1$ ).

Bien que la matrice de covariance des OLS habituelle ne soit généralement pas valable en présence d'hétéroscédasticité, il existe une situation particulière où elle le devient. La différence entre la matrice de covariance des OLS usuelle et la matrice de covariance des OLS généralisés est

$$\sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Omega} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}.$$

<sup>1</sup> Ces résultats ont été obtenus numériquement, à partir de 20.000 simulations. Leur précision doit être exacte au nombre de décimales retenu.

Ici, l'expression clé est le facteur central du second terme, c'est-à-dire  $\mathbf{X}^\top \boldsymbol{\Omega} \mathbf{X}$ . Puisque  $\boldsymbol{\Omega}$  est diagonale, cette matrice est

$$\sum_{t=1}^n \omega_t^2 \mathbf{X}_t^\top \mathbf{X}_t,$$

où  $\mathbf{X}_t$  désigne la  $t^{\text{ième}}$  ligne de  $\mathbf{X}$ . Il s'agit simplement d'une moyenne des matrices  $\mathbf{X}_t^\top \mathbf{X}_t$ , pondérée par les  $\omega_t^2$ . Dans la plupart des cas, ces pondérations seront liées aux lignes correspondantes de la matrice  $\mathbf{X}$ . Cependant, supposons qu'elles ne le soient pas. Alors,

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \omega_t^2 \mathbf{X}_t^\top \mathbf{X}_t \right) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \omega_t^2 \right) \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \mathbf{X}_t^\top \mathbf{X}_t \right). \quad (16.05)$$

Ici, nous avons multiplié chacune des matrices  $\mathbf{X}_t^\top \mathbf{X}_t$  par la limite en probabilité de la pondération moyenne, au lieu de multiplier par les pondérations individuelles. Si les pondérations sont réellement sans rapport avec  $\mathbf{X}_t^\top \mathbf{X}_t$ , c'est une opération tout à fait valable.

Il est clair que l'estimation OLS de la variance d'erreur tendra vers

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \omega_t^2 \right) \equiv \sigma_0^2.$$

Désormais le membre de droite de (16.05) peut se récrire comme

$$\sigma_0^2 \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \mathbf{X}_t^\top \mathbf{X}_t \right) = \sigma_0^2 \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right).$$

Ainsi, si (16.05) est valable, nous pouvons voir que la limite en probabilité de  $n$  fois la matrice de covariance des OLS généralisés (16.03) est

$$\sigma_0^2 \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{X} \right)^{-1}, \quad (16.06)$$

à savoir la matrice de covariance asymptotique conventionnelle des OLS.

On rencontre fréquemment une situation où (16.05) et (16.06) sont vérifiées lorsque  $\mathbf{X}$  contient uniquement un terme constant. Dans ce cas,  $\mathbf{X}_t^\top \mathbf{X}_t$  est égale à 1 pour tout  $t$ , et

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \omega_t^2 \mathbf{X}_t^\top \mathbf{X}_t \right) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \omega_t^2 \right) = \sigma_0^2.$$

Par conséquent, si nous estimons une moyenne, la formule usuelle pour l'écart type d'une moyenne d'échantillon sera valable qu'il y ait hétéroscédasticité ou pas.

Tous les résultats qui précèdent se généralisent facilement au cas de la régression non linéaire. Supposons que nous estimions le modèle de régression non linéaire  $\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}$  par NLS lorsque le DGP est

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0) + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \boldsymbol{\Omega},$$

où  $\boldsymbol{\Omega}$  possède les mêmes propriétés que celles supposées dans le cas linéaire. Alors il n'est pas difficile de voir que la relation asymptotique suivante, équivalente à l'équation (5.39) réécrite d'une manière légèrement différente, est tout aussi valable que dans le cas homoscédastique:

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} (n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1}n^{-1/2}\mathbf{X}_0^\top\mathbf{u}. \quad (16.07)$$

Ici,  $\mathbf{X}_0$  désigne  $\mathbf{X}(\boldsymbol{\beta}_0)$ , la matrice des dérivées de  $\mathbf{x}(\boldsymbol{\beta})$  par rapport à  $\boldsymbol{\beta}$ , évaluée en  $\boldsymbol{\beta}_0$ . A partir de (16.07), nous concluons immédiatement que la matrice de covariance asymptotique de l'estimateur NLS est

$$\text{plim}_{n \rightarrow \infty} \left( (n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1} (n^{-1}\mathbf{X}_0^\top\boldsymbol{\Omega}\mathbf{X}_0) (n^{-1}\mathbf{X}_0^\top\mathbf{X}_0)^{-1} \right). \quad (16.08)$$

C'est, bien sûr, l'analogue immédiat de la seconde ligne de (16.03).

### 16.3 ESTIMATION DE LA MATRICE DE COVARIANCE

A première vue, l'estimateur de la matrice de covariance des OLS généralisés et son analogue NLS (16.08) ne semblent pas très utiles. Il nous faut connaître  $\boldsymbol{\Omega}$  pour les calculer, mais si c'était le cas, nous pourrions utiliser les GLS ou les GNLS et obtenir des estimations plus efficaces. Ce sentiment était dominant chez les économètres jusqu'au début des années 80. Mais, un article très influent de White (1980) a montré qu'il est possible d'obtenir un estimateur de la matrice de covariance des estimations par moindres carrés asymptotiquement valable même en présence d'hétéroscédasticité de forme inconnue.<sup>2</sup> Un tel estimateur est appelé **estimateur de la matrice de covariance robuste à l'hétéroscédasticité**, ou **HCCME**.

L'astuce pour obtenir un HCCME est de reconnaître qu'il *n'est pas* nécessaire d'estimer  $\boldsymbol{\Omega}$  de manière convergente. Cela serait de toute manière une tâche insurmontable puisque  $\boldsymbol{\Omega}$  comporte  $n$  éléments diagonaux à estimer. La matrice de covariance asymptotique d'un vecteur d'estimations NLS, sous l'hypothèse d'hétéroscédasticité, est donnée par l'expression (16.08), qui peut se récrire comme

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}_0^\top \mathbf{X}_0 \right)^{-1} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}_0^\top \boldsymbol{\Omega} \mathbf{X}_0 \right) \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}_0^\top \mathbf{X}_0 \right)^{-1}. \quad (16.09)$$

<sup>2</sup> Les précurseurs de l'article de White dans la littérature statistique sont Eicker (1963, 1967) et Hinkley (1977), ainsi que certains articles novateurs sur le "bootstrap" (consulter le Chapitre 21).



Les premier et troisième facteurs sont ici identiques, et nous pouvons les estimer facilement avec l'approche usuelle. Un estimateur convergent est

$$\frac{1}{n} \hat{\mathbf{X}}^\top \hat{\mathbf{X}},$$

où  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\boldsymbol{\beta}})$ . Alors, la seule difficulté consiste à estimer le second facteur. White a montré que ce second facteur peut être estimé de manière convergente par

$$\frac{1}{n} \hat{\mathbf{X}}^\top \hat{\boldsymbol{\Omega}} \hat{\mathbf{X}}, \quad (16.10)$$

où  $\hat{\boldsymbol{\Omega}}$  peut être un parmi de nombreux estimateurs *non convergents* de  $\boldsymbol{\Omega}$ . La version la plus simple de  $\hat{\boldsymbol{\Omega}}$ , et aussi celle que proposa White dans le contexte des modèles de régression linéaire, a pour  $t^{\text{ième}}$  élément diagonal  $\hat{u}_t^2$ , le  $t^{\text{ième}}$  résidu au carré des moindres carrés.

Contrairement à  $\boldsymbol{\Omega}$ , le facteur central de (16.09) n'a que  $\frac{1}{2}(k^2 + k)$  éléments distincts, quelle que soit la taille de l'échantillon. C'est pourquoi il est possible de l'estimer de façon convergente. Un élément type de cette matrice est

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \omega_t^2 X_{ti} X_{tj} \right), \quad (16.11)$$

où  $X_{ti} \equiv X_{ti}(\boldsymbol{\beta}_0)$ . D'autre part, un élément type de (16.10) est

$$\frac{1}{n} \sum_{t=1}^n \hat{u}_t^2 \hat{X}_{ti} \hat{X}_{tj}. \quad (16.12)$$

Du fait que  $\hat{\boldsymbol{\beta}}$  converge vers  $\boldsymbol{\beta}_0$ ,  $\hat{u}_t$  converge vers  $u_t$ ,  $\hat{u}_t^2$  vers  $u_t^2$ , et  $\hat{X}_{ti}$  converge vers  $X_{ti}$ . Par conséquent, l'expression (16.12) est asymptotiquement égale à

$$\frac{1}{n} \sum_{t=1}^n u_t^2 X_{ti} X_{tj}. \quad (16.13)$$

Sous nos hypothèses, nous pouvons appliquer une loi des grands nombres à (16.13); voir White (1980, 1984) et Nicholls et Pagan (1983) pour certains détails techniques. Il s'ensuit immédiatement que (16.13), et par conséquent (16.12) aussi, tendent en probabilité vers (16.11). Par conséquent, la matrice

$$(n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} (n^{-1} \hat{\mathbf{X}}^\top \hat{\boldsymbol{\Omega}} \hat{\mathbf{X}}) (n^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \quad (16.14)$$

est une estimation convergente de (16.09). Bien entendu, dans la pratique, nous ignorons les facteurs  $n^{-1}$  et utilisons la matrice

$$(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top \hat{\boldsymbol{\Omega}} \hat{\mathbf{X}} (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \quad (16.15)$$

pour estimer la matrice de covariance de  $\hat{\boldsymbol{\beta}}$ .

Des inférences asymptotiquement valables sur  $\beta$  peuvent se baser sur le HCCME (16.15) suivant la procédure habituelle. Cependant, il faut être prudent lorsque  $n$  est petit. Il y a de fortes chances que cet HCCME soit peu fiable en échantillon fini. Après tout, le fait que (16.14) estime (16.09) de manière convergente n'implique pas que (16.14) estime toujours très bien (16.09) en échantillon fini.

Il est possible de modifier le HCCME (16.15) pour lui conférer de meilleures propriétés en échantillon fini. Le problème majeur est que les résidus au carré des moindres carrés ne sont pas des estimations sans biais des aléas au carré  $u_t^2$ . Le moyen le plus simple pour améliorer le HCCME consiste simplement à multiplier (16.15) par  $n/(n-k)$ . Cela revient à diviser la somme des résidus au carré par  $n-k$  plutôt que de diviser par  $n$  pour obtenir l'estimateur de la variance OLS  $s^2$ . Une deuxième approche, bien meilleure, consiste à définir le  $t^{\text{ième}}$  élément diagonal de  $\hat{\Omega}$  par  $\hat{u}_t^2/(1-\hat{h}_t)$ , où  $\hat{h}_t \equiv \hat{\mathbf{X}}_t(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}_t^\top$  est le  $t^{\text{ième}}$  élément diagonal de la matrice “chapeau”  $\hat{\mathbf{P}}_X$  qui projette orthogonalement sur l'espace engendré par les colonnes de  $\hat{\mathbf{X}}$ . Souvenons-nous de la Section 3.2, dans le cas OLS avec une variance constante  $\sigma^2$ , que l'espérance de  $\hat{u}_t^2$  est  $\sigma^2(1-h_t)$ . Par conséquent, dans le cas linéaire, la division de  $\hat{u}_t^2$  par  $1-h_t$  conduirait à une estimation sans biais de  $\sigma^2$  si les aléas étaient effectivement homoscédastiques.

Une troisième possibilité est d'utiliser une technique appelée “jackknife” que nous ne tenterons pas de traiter ici; voir MacKinnon et White (1985). Le HCCME qui en résulte est passablement compliqué mais peut être simplement approximé en définissant le  $t^{\text{ième}}$  élément diagonal de  $\hat{\Omega}$  comme

$$\frac{\hat{u}_t^2}{(1-\hat{h}_t)^2}. \quad (16.16)$$

Il semble que cela puisse induire une correction trop forte de la tendance des résidus des moindres carrés à être trop petits, puisque dans le cas linéaire avec homoscédasticité l'espérance de (16.16) serait plus grande que  $\sigma^2$ . Mais lorsque les aléas manifestent effectivement de l'hétéroscédasticité, les observations à grande variance tendront à influencer les estimation paramétriques plus lourdement que les observations à faible variance et tendront ainsi à avoir des résidus beaucoup trop petits. Ainsi, dans la mesure où de grandes variances sont associées à de grandes valeurs de  $\hat{h}_t$ , cette correction trop forte peut être effectivement un atout.

Nous avons mentionné quatre HCCME différents. Nous les noterons de  $HC_0$  à  $HC_3$ . C'est leur définition du  $t^{\text{ième}}$  élément de  $\hat{\Omega}$  qui les distingue:

$$\begin{aligned} HC_0: & \hat{u}_t^2 \\ HC_1: & \frac{n}{n-k} \hat{u}_t^2 \end{aligned}$$

$$HC_2: \frac{\hat{u}_t^2}{1 - \hat{h}_t}$$

$$HC_3: \frac{\hat{u}_t^2}{(1 - \hat{h}_t)^2}.$$

MacKinnon et White (1985) ont étudié le fonctionnement en échantillon fini des statistiques pseudo- $t$  basées sur ces quatre HCCME.<sup>3</sup> Ils ont trouvé tout d'abord que  $HC_0$  possède les moins bonnes performances, tendant à rejeter beaucoup trop fréquemment l'hypothèse nulle dans certains cas, que  $HC_1$  fonctionnait mieux et  $HC_2$  encore mieux, qu'enfin  $HC_3$  possède les meilleures propriétés. Un travail ultérieur de Chesher et Jewitt (1987), Chesher (1989), et Chesher et Austin (1991) a ajouté une certaine précision aux fondements de ces résultats et a montré que  $HC_3$  ne fonctionnera pas toujours mieux que  $HC_2$ .

Pour une raison pratique, nous ne devrions jamais employer  $HC_0$ , puisque  $HC_1$  n'est pas plus coûteux en temps de calcul et fonctionne toujours mieux. Lorsque les éléments diagonaux de la matrice "chapeau" sont disponibles, nous devrions utiliser sans hésitation  $HC_2$  ou  $HC_3$  plutôt que  $HC_1$ . Cependant, le choix de la forme à employer n'est pas tout à fait clair. Dans certains cas,  $HC_2$  est plus attrayant, mais  $HC_3$  semble généralement mieux fonctionner dans les expériences Monte Carlo.

Bien que de nombreux progiciels de régression calculent maintenant les HCCME, ils nous fournissent souvent le HCCME le moins souhaitable, c'est-à-dire  $HC_0$ . Messer et White (1984) suggèrent un moyen ingénieux de calculer n'importe lequel de ces HCCME grâce à un programme conçu pour l'estimation par variables instrumentales. Notons  $\tilde{u}_t$  une estimation quelconque de  $u_t$ :  $\hat{u}_t$  dans le cas de  $HC_0$ ,  $\hat{u}_t/(1 - \hat{h}_t)^{1/2}$  dans le cas de  $HC_2$ , et ainsi de suite. La procédure suggérée par Messer et White consiste à construire les variables artificielles

$$y_t^* \equiv \frac{y_t}{\tilde{u}_t}, \quad \mathbf{X}_t^* \equiv \frac{\mathbf{X}_t}{\tilde{u}_t}, \quad \text{et} \quad \mathbf{Z}_t \equiv \mathbf{X}_t \tilde{u}_t$$

et à régresser  $y_t^*$  sur  $\mathbf{X}_t^*$  en se servant de  $\mathbf{Z}_t$  comme d'un vecteur d'instruments. Les estimations IV obtenues de cette façon sont identiques à celles d'une régression OLS de  $y_t$  sur  $\mathbf{X}_t$ , et la matrice de covariance IV est proportionnelle au HCCME correspondant à n'importe quel ensemble de résidus  $\tilde{u}_t$  retenu. Le facteur de proportionnalité est  $s^2$ , à savoir l'estimation IV de la variance d'erreur, qui tendra vers 1 lorsque  $n \rightarrow \infty$ . Ainsi, à moins que  $s^2 = 1$ , ce qui sera le cas si  $\tilde{u}_t = \hat{u}_t$  et le progiciel de régression divisera par  $n$  plutôt que par  $n - k$ , nous divisons simplement la matrice de covariance IV par  $s^2$ .

<sup>3</sup> En réalité, ils ont recherché le fonctionnement du "jackknife" plutôt que du HCCME appelé  $HC_3$  mais d'autres simulations de calculs suggèrent que  $HC_3$  se comporte de façon similaire au "jackknife".

Cette procédure n'est tolérée que si aucun des  $\tilde{u}_t$  n'est identiquement nul; les manières de gérer des résidus nuls sont discutées dans l'article d'origine. Bien entendu, n'importe quel HCCME peut être calculé directement en utilisant de nombreux langages de programmation différents. L'élément clé consiste à lancer les calculs de telle sorte que la matrice  $\hat{\Omega}$  de dimension  $n \times n$  ne soit jamais formée explicitement.

Il existe deux moyens distincts d'employer les HCCME pour tester les hypothèses. Le plus direct consiste simplement à construire des tests de Wald et des statistiques pseudo- $t$ , à la manière habituelle, en utilisant le HCCME à la place de l'estimateur usuel de la matrice de covariance des moindres carrés. Cependant, comme nous l'avons vu dans la Section 11.6, il est également possible de construire des tests LM, ou  $C(\alpha)$ , basés sur ce que nous avons appelé régression de Gauss-Newton robuste à l'hétéroscédasticité, ou HRGNR. Supposons que l'hypothèse alternative soit

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{u},$$

avec  $\boldsymbol{\beta}$  un vecteur de dimension  $k$  et  $\boldsymbol{\gamma}$  un vecteur de dimension  $r$ , où l'hypothèse nulle est  $\boldsymbol{\gamma} = \mathbf{0}$ . Notons  $\dot{\mathbf{X}}$  et  $\dot{\mathbf{Z}}$  les matrices des dérivées partielles de  $\mathbf{x}(\boldsymbol{\beta}, \boldsymbol{\gamma})$  par rapport aux éléments de  $\boldsymbol{\beta}$  et  $\boldsymbol{\gamma}$  respectivement, évaluées avec les estimations convergentes au taux  $n^{1/2}$   $[\dot{\boldsymbol{\beta}}; \mathbf{0}]$ . Alors, si  $\dot{\mathbf{M}}_X$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\dot{\mathbf{X}})$ ,  $\dot{\mathbf{u}}$  désigne un vecteur de résidus de dimension  $n$  d'élément type  $\dot{u}_t = y_t - x_t(\dot{\boldsymbol{\beta}}, \mathbf{0})$ , et  $\dot{\Omega}$  désigne une matrice diagonale de dimension  $n \times n$  d'élément diagonal type  $\dot{u}_t^2$ , la statistique de test

$$\dot{\mathbf{u}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{Z}} (\dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\Omega} \dot{\mathbf{M}}_X \dot{\mathbf{Z}})^{-1} \dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{u}} \quad (16.17)$$

est asymptotiquement distribuée suivant une  $\chi^2(r)$ . Cette statistique de test est égale à  $n$  moins la somme des résidus au carré de la régression artificielle

$$\boldsymbol{\iota} = \dot{\mathbf{U}} \dot{\mathbf{M}}_X \dot{\mathbf{Z}} \mathbf{b} + \text{résidus},$$

où  $\boldsymbol{\iota}$  est, comme d'habitude, un vecteur à  $n$  composantes où chaque élément est égal à 1, et  $\dot{\mathbf{U}}$  est une matrice diagonale de dimension  $n \times n$  avec le vecteur  $\dot{\mathbf{u}}$  sur la diagonale principale. Nous avons fourni des instructions précises pour calculer (16.17) dans la Section 11.6.

A présent, il est possible de comprendre précisément comment la HRGNR se déroule. La matrice centrale dans (16.17) est simplement un HCCME pour la matrice de covariance du vecteur  $\dot{\mathbf{Z}}^\top \dot{\mathbf{M}}_X \dot{\mathbf{u}}$ , qui devrait être asymptotiquement d'espérance nulle, si l'hypothèse nulle est correcte. Pour en savoir plus sur la HRGNR, consulter Davidson et MacKinnon (1985b) et Wooldridge (1990a, 1990b, 1991a). Il paraît que la statistique de test de la HRGNR (16.17) est souvent plus proche de sa distribution asymptotique en échantillon fini que ne le sont les statistiques de test de Wald basées sur des HCCME; voir Davidson et MacKinnon (1985b).

L'élément clé sous-jacent au HCCME est que nous pouvons estimer de manière convergente une matrice comme la matrice centrale de (16.09) sans pour autant être capable d'estimer la matrice  $\mathbf{\Omega}$  de manière convergente. Cette idée fondamentale surgira à nouveau dans le prochain chapitre lorsque nous discuterons de la technique d'estimation connue sous le nom de la méthode des moments généralisée. Entre autres, cette technique nous permettra de calculer des estimations asymptotiquement plus efficaces que les estimations par moindres carrés en présence d'une forme inconnue d'hétéroscédasticité.

## 16.4 HÉTÉROSCÉDASTICITÉ DITE ARCH

Les économètres rencontrent fréquemment des modèles estimés à l'aide de séries temporelles où les résidus sont relativement faibles pendant un certain nombre de périodes successives, puis beaucoup plus grands pour un certain nombre d'autres périodes et encore relativement faibles pour une troisième catégorie de périodes et ainsi de suite, et ce généralement sans aucune raison apparente. Ce phénomène est particulièrement fréquent et visible avec des données boursières, des taux de changes étrangers, ou d'autres prix déterminés sur les marchés financiers, où la volatilité semble généralement varier dans le temps. Récemment, d'importants approfondissements ont vu le jour dans la littérature pour modéliser ce phénomène. L'article novateur de Engle (1982b), expose pour la première fois le concept d'**hétéroscédasticité conditionnelle autorégressive**, ou **ARCH**. L'idée fondamentale de l'ARCH est que la variance de l'aléa au temps  $t$  dépend de l'importance des aléas au carré des périodes précédentes. Cependant, il existe plusieurs façons de modéliser cette idée de base, et la littérature correspondante est assez foisonnante.

Notons  $u_t$  le  $t^{\text{ième}}$  aléa associé à un modèle de régression quelconque. Alors, le modèle originel ARCH peut s'écrire

$$\sigma_t^2 \equiv E(u_t^2 | \Omega_t) = \alpha + \gamma_1 u_{t-1}^2 + \gamma_2 u_{t-2}^2 + \cdots + \gamma_p u_{t-p}^2, \quad (16.18)$$

où  $\Omega_t$  désigne l'ensemble d'information sur lequel  $\sigma_t^2$ , la variance de  $u_t$ , doit être conditionnée. Typiquement, cet ensemble d'information contient tous les éléments indicés par  $t - 1$  et par les périodes précédentes. Ce modèle particulier est appelé processus **ARCH(p)**. Sa ressemblance avec le processus **AR(p)** traité dans le Chapitre 10 est frappante et justifie le nom donné à ces modèles. Nous pouvons voir à partir de (16.18) que la variance conditionnelle de  $u_t$  dépend des valeurs de  $u_t^2$  réalisées dans le passé. Pour garantir que cette variance conditionnelle soit toujours positive, nous devons supposer que  $\alpha$  et tous les  $\gamma_i$  ne sont pas négatifs.

La version la plus simple de (16.18) est le processus **ARCH(1)**,

$$\sigma_t^2 = \alpha + \gamma_1 u_{t-1}^2. \quad (16.19)$$

La variance conditionnelle de  $u_t$  obtenue par (16.19) peut être comparée à la variance non conditionnelle  $\sigma^2 \equiv E(u_t^2)$ . En supposant que le processus ARCH(1) est stationnaire, ce qui sera le cas si  $\gamma_1 < 1$ , nous pouvons écrire

$$\sigma^2 = \alpha + \gamma_1 \sigma^2.$$

Cela implique que

$$\sigma^2 = \frac{\alpha}{1 - \gamma_1}.$$

Par conséquent, la variance non conditionnelle de  $u_t$  dépend des paramètres du processus ARCH et sera, en général, différente de la variance conditionnelle donnée par l'équation (16.19).

Sous l'hypothèse nulle d'erreurs homoscedastiques, tous les  $\gamma_i$  sont nuls. Comme Engle (1982b) le montra le premier, il est facile de tester cette hypothèse en exécutant la régression

$$\hat{u}_t^2 = a + c_1 \hat{u}_{t-1}^2 + c_2 \hat{u}_{t-2}^2 + \cdots + c_p \hat{u}_{t-p}^2 + \text{résidu}, \quad (16.20)$$

où  $\hat{u}_t$  désigne un résidu provenant de l'estimation par moindres carrés du modèle de régression auquel sont associés les  $u_t$ . Ensuite, nous calculons un test en  $F$  ordinaire (ou simplement  $n$  fois le  $R^2$  centré) pour les hypothèses de nullité des paramètres  $c_1$  à  $c_p$ . Cette régression artificielle a la même forme que celle utilisée pour tester l'hypothèse d'homoscédasticité discutée dans la Section 11.5. Mais à présent les régresseurs sont des résidus au carré retardés plutôt que des variables indépendantes. Ainsi, pour tout modèle de régression estimé à l'aide de séries temporelles, il est très facile de tester l'hypothèse nulle d'homoscédasticité contre l'hypothèse alternative que les erreurs suivent un processus ARCH( $p$ ).

Le test des erreurs suivant un processus ARCH joue le même rôle dans l'analyse des seconds moments d'un modèle de régression temporelle que le test des erreurs suivant un processus AR dans l'analyse des premiers moments. Tout comme l'évidence d'erreurs AR peut ou pas indiquer que les aléas suivent effectivement un processus AR, l'évidence d'erreurs ARCH peut ou pas à son tour indiquer la présence d'une hétéroscédasticité conditionnelle autorégressive. Dans les deux cas, d'autres formes de mauvaises spécifications peuvent nous conduire sur la piste de ce qui ressemble à des erreurs AR ou ARCH. A partir de légères modifications, l'analyse du Chapitre 12 s'applique aussi bien aux tests dans des directions scédastiques (par exemple, des tests d'hétéroscédasticité) qu'à n'importe quel autre test de spécification.

Plusieurs variantes du modèle ARCH ont été proposées. Une variante particulièrement utile est le modèle **ARCH généralisée**, ou **GARCH**, suggéré par Bollerslev (1986). Le modèle **GARCH**( $p, q$ ) peut s'écrire

$$\sigma_t^2 = \alpha + \sum_{i=1}^p \gamma_i u_{t-i}^2 + \sum_{j=1}^q \delta_j \sigma_{t-j}^2$$

ou, dans une notation plus compacte,

$$\sigma_t^2 = \alpha + A(L, \gamma)u_t^2 + B(L, \delta)\sigma_t^2,$$

où  $\gamma$  et  $\delta$  sont des vecteurs paramétriques d'éléments types  $\gamma_i$  et  $\delta_j$  respectivement, et  $A(L, \gamma)$  et  $B(L, \delta)$  des polynômes de l'opérateur retard  $L$ . Dans le modèle GARCH, la variance conditionnelle  $\sigma_t^2$  dépend aussi bien de ses propres valeurs passées que des valeurs retardées de  $u_t^2$ . Cela signifie que  $\sigma_t^2$  dépend effectivement de toutes les valeurs passées de  $u_t^2$ . Dans la pratique, un modèle GARCH avec très peu de paramètres s'ajuste souvent aussi bien qu'un modèle ARCH ayant de nombreux paramètres. En particulier, un modèle simple qui fonctionne souvent très bien est le modèle **GARCH(1,1)**

$$\sigma_t^2 = \alpha + \gamma_1 u_{t-1}^2 + \delta_1 \sigma_{t-1}^2. \quad (16.21)$$

Dans la pratique, nous devons résoudre un modèle GARCH pour éliminer les termes  $\sigma_{t-j}^2$  de l'expression de droite avant de pouvoir l'estimer. Le problème ressemble essentiellement à l'estimation d'un modèle à moyenne mobile ou d'un modèle ARMA avec une composante moyenne mobile; voir la Section 10.7. Par exemple, le modèle GARCH(1,1) de l'expression (16.21) peut être résolu récursivement pour donner

$$\sigma_t^2 = \frac{\alpha}{1 - \delta_1} + \gamma_1 (u_{t-1}^2 + \delta_1 u_{t-2}^2 + \delta_1^2 u_{t-3}^2 + \delta_1^3 u_{t-4}^2 + \cdots). \quad (16.22)$$

Différentes hypothèses peuvent être retenues concernant des aléas antérieurs à ceux de l'échantillon. La plus simple est de supposer qu'ils sont nuls, mais il est plus réaliste de supposer qu'ils sont égaux à leur espérance non conditionnelle.

Il est intéressant d'observer que lorsque  $\delta_1$  est proche de zéro, le modèle résolu GARCH(1,1) (16.22) ressemble à un modèle ARCH(2). Pour cette raison, il s'avère qu'un test approprié pour des erreurs GARCH(1,1) doit simplement régresser les résidus au carré sur un terme constant et les résidus au carré retardés à l'ordre un et deux. En général, un test LM d'erreurs GARCH( $p, q$ ) est équivalent à un test LM d'erreurs ARCH(max( $p, q$ )). Ces résultats sont totalement analogues à ceux des tests d'erreurs ARMA( $p, q$ ) discutés dans la Section 10.8.

Il existe trois principaux moyens d'estimer des modèles de régression à erreurs ARCH et GARCH: les GLS faisables, l'estimation efficace en une étape, et le maximum de vraisemblance. Dans l'approche la plus simple, celle des GLS faisables, nous estimons tout d'abord le modèle de régression par moindres carrés ordinaires ou non linéaires, puis utilisons les résidus au carré pour estimer les paramètres d'un processus ARCH ou GARCH, et enfin utilisons les moindres carrés pondérés pour estimer les paramètres de la fonction de régression. Cette procédure peut se heurter à des difficultés si les variances conditionnelles prédites par le processus ajusté ARCH ne sont pas toutes positives, et de nombreuses autres méthodes *ad hoc* peuvent alors être utilisées pour garantir qu'elles sont toutes positives.

Les estimations des paramètres ARCH obtenues à l'aide de cette variante de GLS faisables ne seront pas asymptotiquement convergentes. C'est pour cela que Engle (1982b) suggéra d'utiliser une version de l'estimation efficace en une étape. Quoiqu'il en soit, cette méthode est un peu trop compliquée pour être discutée ici.

La troisième méthode d'estimation largement répandue consiste à utiliser le maximum de vraisemblance en supposant la normalité des aléas. Supposons que le modèle à estimer soit un modèle de régression non linéaire dont les aléas GARCH( $p, q$ ) sont conditionnellement normaux:

$$\begin{aligned} y_t &= x_t(\beta) + u_t, & u_t &= \sigma_t \varepsilon_t, \\ \sigma_t^2 &= \alpha + A(L, \gamma)u_t^2 + B(L, \delta)\sigma_t^2, & \varepsilon_t &\sim \text{NID}(0, 1). \end{aligned} \quad (16.23)$$

La fonction de logvraisemblance pour ce modèle est

$$C - \frac{1}{2} \sum_{t=1}^n \log(\sigma_t^2(\alpha, \gamma, \delta, \beta)) - \frac{1}{2} \sum_{t=1}^n \frac{(y_t - x_t(\beta))^2}{\sigma_t^2(\alpha, \gamma, \delta, \beta)}, \quad (16.24)$$

où  $C$  est une constante et

$$\sigma_t^2(\alpha, \gamma, \delta, \beta) \equiv \alpha + A(L, \gamma)(y_t - x_t(\beta))^2 + B(L, \delta)\sigma_t^2. \quad (16.25)$$

Comme il s'agit d'un modèle GARCH, nous devons déterminer  $\sigma_t^2$  à partir de (16.25) de manière récursive, pour pouvoir évaluer (16.24). L'algèbre est assez compliquée, mais avec un progiciel approprié, l'estimation n'est pas trop difficile.

Le modèle (16.23) est clairement un modèle pour lequel la régression artificielle à longueur double (DLR), introduite dans la Section 14.4, est applicable. Si nous prenons la définition

$$f_t(y_t, \theta) \equiv \frac{y_t - x_t(\beta)}{(\alpha + A(L, \gamma)(y_t - x_t(\beta))^2 + B(L, \delta)\sigma_t^2)^{1/2}},$$

il est clair que ce modèle est un cas particulier de la classe des modèles (14.18). L'obtention des dérivées partielles nécessaires à l'exécution de la DLR n'est pas triviale, tout particulièrement lorsque  $\delta \neq \mathbf{0}$ , mais le même effort est également nécessaire pour exécuter n'importe quelle technique d'estimation asymptotiquement convergente. La DLR peut être utilisée pour obtenir des estimations efficaces en une étape, à partir d'estimations OLS et d'estimations convergentes des paramètres ARCH obtenues, ou en tant que composante d'une procédure d'estimation ML. Bien sûr, la DLR nous fournit aussi un moyen naturel et relativement commode de calculer une large variété de tests de spécification pour les modèles à erreurs ARCH et GARCH.



Un des nombreux développements autour du concept originel ARCH se trouve dans la classe importante de modèles appelée classe **ARCH-en-moyenne**, ou **ARCH-M**. Cette classe de modèles a été introduite par Engle, Lilien, et Robins (1987). Ces modèles sont comparables aux autres modèles ARCH, excepté que la variance conditionnelle  $\sigma_t^2$  intervient dans la fonction de régression pour conditionner l'espérance. Ainsi (16.23) deviendrait

$$y_t = x_t(\beta, \sigma_t^2) + u_t, \quad u_t = \sigma_t \varepsilon_t, \\ \sigma_t^2 = \alpha + A(L, \gamma)u_t^2 + B(L, \delta)\sigma_t^2, \quad \varepsilon_t \sim \text{NID}(0, 1).$$

De nombreuses théories en économie financière utilisent des mesures du risque. Si nous assimilons la variance conditionnelle d'un aléa à une mesure du risque, il semble logique que  $\sigma_t^2$  puisse intervenir dans la fonction de régression en tant que mesure du risque. La matrice d'information des modèles ARCH-M n'est pas bloc-diagonale entre  $\beta$  d'une part et  $\gamma$  et  $\delta$  d'autre part. Par conséquent, les GLS faisables et les techniques d'estimation efficace en une étape qui s'appliquent à d'autres modèles ARCH ne peuvent pas être utilisées. Le maximum de vraisemblance est la technique qui est presque toujours employée.

La littérature sur la modélisation ARCH est assez conséquente et ne cesse de croître. Engle et Bollerslev (1986) fournissent un article de synthèse très utile sur les premiers travaux dans ce domaine. Engle et Rothschild (1992) proposent un éventail d'articles récents, incluant Bollerslev, Chou, et Kroner (1992), dont les références bibliographiques sont très intéressantes. Engle, Hendry, et Trumble (1985) proposent une démonstration à l'aide d'expériences Monte Carlo sur les propriétés des estimateurs ARCH et des statistiques de test en échantillon fini. Des articles plus appliqués furent écrits par Domowitz et Hakkio (1985), Bollerslev, Engle, et Wooldridge (1988), McCurdy et Morgan (1988), et Nelson (1991).

## 16.5 TESTS D'HÉTÉROSCÉDASTICITÉ

Dans la Section 11.5, nous avons discuté de plusieurs tests d'hétéroscédasticité basés sur des régressions artificielles comparables à la régression de Gauss-Newton, dans lesquelles la régressande était un vecteur de résidus au carré. Dans cette section, nous discutons plus en détails de ces tests, ainsi que d'autres tests d'hétéroscédasticité.

Supposons que l'hypothèse nulle soit

$$y_t = x_t(\beta) + u_t, \quad u_t \sim \text{IID}(0, \sigma^2),$$

et l'hypothèse alternative que la fonction de régression soit encore  $x_t(\beta)$ , mais avec

$$E(u_t^2) = h(\alpha + \mathbf{Z}_t\gamma),$$

où  $h(\cdot)$  est une fonction prenant des valeurs positives qui peut être linéaire ou non,  $\mathbf{Z}_t$  un vecteur de  $q$  observations sur des variables exogènes ou prédéterminées, et  $\alpha$  et  $\gamma$  des paramètres inconnus. Nous avons vu dans la Section 11.5 que l'on peut tester l'hypothèse  $\gamma = \mathbf{0}$  en testant  $\mathbf{c} = \mathbf{0}$  dans la régression artificielle

$$\hat{\mathbf{v}} = \boldsymbol{\iota}a^* + \mathbf{Z}\mathbf{c} + \text{résidus}. \quad (16.26)$$

Ici,  $\hat{\mathbf{v}}$  désigne un vecteur d'élément type  $\hat{u}_t^2$ ,  $\boldsymbol{\iota}$  un vecteur dont chaque élément égale 1, et  $\mathbf{Z}$  une matrice de ligne type  $\mathbf{Z}_t$ . La statistique de test est soit  $n$  fois le  $R^2$  centré, soit la statistique ordinaire en  $F$  pour  $\mathbf{c} = \mathbf{0}$ . Nous avons dérivé ce test comme une application des résultats généraux sur les régressions de Gauss-Newton.

Dans la Section 11.5, nous nous sommes peu étendus sur la manière de choisir la matrice  $\mathbf{Z}$ . Il existe un grand nombre de moyens d'y parvenir. Cette matrice  $\mathbf{Z}$  peut se composer de n'importe quelle observation sur des variables exogènes ou prédéterminées appartenant à l'ensemble d'information qui conditionne  $\mathbf{y}$ , ou des fonctions de telles variables, et elle peut avoir une ou plusieurs colonnes. Une approche consiste à spécifier des alternatives hétéroscédastiques particulières qui semblent plausibles et à dériver la matrice  $\mathbf{Z}$  en conséquence. La régression (16.20) utilisée pour tester des erreurs ARCH( $p$ ) est un exemple particulier dans lequel la matrice  $\mathbf{Z}$  est composée exclusivement de résidus au carré retardés. Un autre exemple est celui de l'hétéroscédasticité multiplicative, qui semble souvent plausible si la régressande est toujours largement supérieure à zéro. Par conséquent, dans ce cas, une hypothèse alternative raisonnable serait

$$E(u_t^2) = \alpha(\mathbf{X}_t\boldsymbol{\beta})^\gamma. \quad (16.27)$$

Puisque l'hypothèse nulle correspond à  $\gamma = 0$ , on peut assimiler  $\alpha$  à  $\sigma^2$ . La dérivée partielle de la partie droite de (16.27) par rapport à  $\gamma$  est

$$\alpha(\mathbf{X}_t\boldsymbol{\beta})^\gamma \log(\mathbf{X}_t\boldsymbol{\beta}). \quad (16.28)$$

L'évaluation de (16.28) sous l'hypothèse nulle que  $\gamma = 0$  conduit à une expression simplifiée  $\hat{\sigma}^2 \log(\mathbf{X}_t\hat{\boldsymbol{\beta}})$ . Par conséquent, pour tester l'hypothèse  $\gamma = 0$ , nous devons simplement régresser  $\hat{u}_t^2$  sur une constante et  $\log(\mathbf{X}_t\hat{\boldsymbol{\beta}})$ . La statistique de test est le  $t$  de Student de  $\log(\mathbf{X}_t\hat{\boldsymbol{\beta}})$ . Elle devrait être distribuée asymptotiquement selon une  $N(0, 1)$  sous l'hypothèse nulle.

Plusieurs spécifications de l'hétéroscédasticité peuvent, comme (16.27), sembler plausibles dans des cas particuliers. Celles-ci conduisent à différentes spécifications de la matrice  $\mathbf{Z}$ . Quand il y a une bonne raison *a priori* de supposer que l'hétéroscédasticité prend une forme particulière, il est pertinent de tester cette forme, et d'utiliser alors les GLS faisables ou le ML pour la prendre en compte si l'hypothèse nulle est rejetée.

Par ailleurs, s'il existe, *a priori*, peu d'information concernant la forme que peut prendre l'hétéroscédasticité si elle est présente, la spécification de la matrice  $\mathbf{Z}$  devient beaucoup plus difficile. Une approche fut suggérée par White (1980). Nous avons vu dans la Section 16.2, pour un modèle de régression linéaire estimé par OLS, que la matrice de covariance OLS conventionnelle est asymptotiquement valable pourvu que l'espérance non conditionnelle  $E(u_t^2)$  soit égale à l'espérance conditionnelle aux carrés et aux produits croisés de tous les régresseurs. Par conséquent, White suggéra que la matrice  $\mathbf{Z}$  devrait se composer des carrés et des produits croisés de tous les régresseurs, excepté le terme constant et toutes les autres colonnes qui empêcheraient  $[\mathbf{I} \ \mathbf{Z}]$  d'être de plein rang. Comme tous les tests basés sur des régressions que nous avons abordés, le **test de White** aura un paramètre de non centralité différent de zéro chaque fois qu'il y a une quelconque corrélation entre  $u_t^2$  et n'importe lequel des éléments de  $\mathbf{Z}_t$ . Ainsi, si l'échantillon est assez grand, le test de White est certain de détecter n'importe quelle hétéroscédasticité qui provoquerait la non convergence de la matrice de covariance OLS.

Bien que le test de White soit convergent pour un éventail d'alternatives d'hétéroscédasticité très large, il peut ne pas être très puissant en échantillon fini. Le problème est que le nombre de colonnes dans la matrice  $\mathbf{Z}$  sera très grand si le nombre de régresseurs dans la matrice  $\mathbf{X}$  n'est pas suffisamment restreint. En général, la matrice  $\mathbf{Z}$  aura  $k(k+1)/2 - 1$  colonnes si  $\mathbf{X}$  comprend un terme constant. Ainsi, pour  $k = 10$ , le test de White aura 54 degrés de liberté; pour  $k = 20$  (ce qui n'est finalement pas très grand pour des études utilisant des données en coupe transversale), il y aura 209 degrés de liberté. Ce sont des nombres plutôt grands. Comme nous l'avons vu dans le Chapitre 12, les tests possédant des degrés de liberté nombreux manqueront probablement de puissance à moins que la taille de l'échantillon ne soit très grande. Nous pourrions, bien entendu, modifier le test de White par différents moyens *ad hoc*, notamment en éliminant les colonnes de la matrice  $\mathbf{Z}$  qui correspondent aux produits croisés des régresseurs. De telles modifications pourraient ou pas améliorer la puissance du test. Cela dépendrait beaucoup de l'importance de la réduction du paramètre de non centralité, comparativement à l'effet de quelques degrés de liberté en moins; consulter la Section 12.5.

Sous l'hypothèse de normalité des aléas, il est possible de formuler des tests LM pour l'hétéroscédasticité en termes de régressions artificielles. Il s'agit en fait de la façon dont ils furent initialement obtenus par Godfrey (1978c) et Breusch et Pagan (1979). L'hypothèse maintenue est

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, h(\alpha + \mathbf{Z}_t\boldsymbol{\gamma})), \quad (16.29)$$

et nous souhaitons tester la contrainte  $\boldsymbol{\gamma} = \mathbf{0}$ . Le test LM peut être dérivé de façon assez immédiate en écrivant la fonction de logvraisemblance correspondant à (16.29), obtenant ainsi le gradient et la matrice d'information, et en évaluant ces deux quantités avec les estimations NLS de  $\boldsymbol{\beta}$ , en construisant enfin la forme quadratique usuelle. Nous laissons ceci en exercice pour le lecteur.

La statistique de test LM que nous obtenons ainsi peut s'écrire comme

$$\frac{1}{2\hat{\sigma}^4} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \mathbf{Z} (\mathbf{Z}^\top \mathbf{M}_\iota \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{M}_\iota \hat{\mathbf{v}}, \quad (16.30)$$

où, comme auparavant,  $\hat{\mathbf{v}}$  est un vecteur à  $n$  composantes, d'élément type  $\hat{u}_t^2$ , et  $\mathbf{M}_\iota$  est la matrice qui calcule les écarts à la moyenne. Sous l'hypothèse nulle, cette statistique de test sera asymptotiquement distribuée selon une  $\chi^2(q)$ . L'expression (16.30) est égale à la moitié de la somme des carrés expliqués de la régression

$$\frac{\hat{\mathbf{v}}}{\hat{\sigma}^2} - \boldsymbol{\iota} = a\boldsymbol{\iota} + \mathbf{Z}\mathbf{c} + \text{résidus}. \quad (16.31)$$

Ce résultat provient du fait que la régressande est ici d'espérance nulle par construction. Le Théorème FWL implique alors que la somme des carrés expliqués de (16.31) reste inchangée si tous les régresseurs sont remplacés par leurs écarts à leur moyenne.

Il est facile de voir pourquoi (16.30) doit être asymptotiquement distribuée selon une  $\chi^2(q)$ . L'hypothèse de normalité implique que  $\hat{u}_t^2/\hat{\sigma}^2$  est distribuée asymptotiquement selon une  $\chi^2(1)$ . Le vecteur

$$n^{-1/2}(\hat{\mathbf{v}}/\hat{\sigma}^2)^\top \mathbf{M}_\iota \mathbf{Z} \quad (16.32)$$

est simplement une moyenne pondérée de  $n$  variables aléatoires, chacune suivant initialement une  $\chi^2(1)$ , mais étant recentrée pour avoir une espérance nulle du fait de la présence de  $\mathbf{M}_\iota$ . Puisque la variance d'une variable aléatoire suivant une distribution  $\chi^2(1)$  est égale à 2 (voir la Section 4 de l'Annexe B), le vecteur (16.32) doit avoir la matrice de covariance

$$\frac{2}{n} \mathbf{Z}^\top \mathbf{M}_\iota \mathbf{Z}. \quad (16.33)$$

La statistique de test LM (16.30) est simplement une forme quadratique du vecteur (16.32) et de l'inverse de la matrice (16.33). A condition qu'un théorème de la limite centrale s'applique à (16.32), sous des conditions assez faibles portant sur la matrice  $\mathbf{Z}$ , cette forme quadratique doit avoir asymptotiquement la distribution  $\chi^2(q)$ .

Le test LM (16.30) est très étroitement relié aux tests  $nR^2$  et en  $F$  basés sur la régression (16.26) déjà analysée. En fait, les  $R^2$  centrés issus de (16.26) et (16.31) sont numériquement identiques, puisque la seule différence entre ces deux régressions est que la régressande de (16.31) a été réduite, et translatée de telle sorte qu'elle ait une moyenne empirique nulle. Le résultat que la statistique LM égale la moitié de la somme des carrés expliqués de (16.31) dépend crucialement de l'hypothèse de normalité, supposée pour calculer (16.33). Sans cette hypothèse, dont Koenker (1981) et d'autres ont critiquée la fiabilité dans la plupart des cas, il ne nous resterait qu'un test en  $F$  ou  $nR^2$ , comme auparavant.

La statistique de test basée sur l'hypothèse de normalité s'avère quelque peu plus puissante que les tests en  $F$  et  $nR^2$  calculés à partir de la même régression artificielle, parce que l'hétéroscédasticité crée souvent l'apparence d'un excès de kurtosis, qui tend à réduire la valeur de n'importe quel test qui utilise une estimation de la variance de  $\hat{u}_t^2$ . Un autre avantage du test LM (16.30) est qu'il peut être modifié pour être presque exact en échantillon fini; voir Honda (1988). Ainsi, cette forme du test LM bénéficie de nombreux avantages, si l'hypothèse de normalité est raisonnable. Comme cette hypothèse peut être très facilement testée, comme nous le verrons dans la Section 16.7, il peut être raisonnable d'utiliser les tests LM pour l'hétéroscédasticité lorsque l'hypothèse de normalité paraît convenir aux données.

Jusqu'à présent, notre discussion des tests d'hétéroscédasticité était exclusivement focalisée sur des tests basés sur des régressions artificielles. De nombreux autres tests furent proposés, et certains d'entre eux sont largement exploités. Un test particulièrement connu est le vénérable test en  $F$  de Goldfeld et Quandt (1965), facile à calculer et souvent très performant. L'idée consiste à ranger les données selon la valeur d'une variable quelconque censée être responsable de l'hétéroscédasticité, puis à estimer le modèle sur les premier et dernier tiers de l'échantillon, puis à calculer la statistique de test

$$\frac{SSR_3/(n_3 - k)}{SSR_1/(n_1 - k)}, \quad (16.34)$$

où  $SSR_1$  et  $SSR_3$  désignent les sommes des résidus au carré des premier et dernier tiers de l'échantillon, et  $n_1$  et  $n_3$  les tailles de l'échantillon associées. En supposant que les aléas sont distribués selon une loi normale, cette statistique de test serait, sous l'hypothèse nulle, exactement distribuée selon une  $F(n_3 - k, n_1 - k)$ ; même sans normalité, elle aurait approximativement cette distribution en grands échantillons. Notons que le **test de Goldfeld-Quandt** est un test bilatéral, caractère plutôt inhabituel pour un test en  $F$ , puisque nous voulons rejeter l'hypothèse nulle si la statistique de test (16.34) est soit trop grande, soit trop petite.

D'autres tests pour l'hétéroscédasticité intéressants furent proposés par Glejser (1969), Szroeter (1978), Harrison et McCabe (1979), Ali et Giacotto (1984), Evans et King (1985, 1988), et Newey et Powell (1987). Plusieurs de ces articles, et notamment ceux de MacKinnon et White (1985) et Griffiths et Surekha (1986), fournissent des démonstrations issues d'expériences Monte Carlo sur les propriétés d'un ou plusieurs tests. Voir aussi Godfrey (1988, Sections 4.5 et 5.5).

## 16.6 DIRECTIONS SCÉDASTIQUES ET DE RÉGRESSION

Dans le Chapitre 12, nous avons présenté une analyse relativement détaillée de ce qui détermine la puissance des tests orientés régression, c'est-à-dire

des tests qui permettent de conclure si la fonction de régression est correctement spécifiée ou non. Une analyse similaire pourrait être entreprise sur la puissance des tests dans des **directions scédastiques**, c'est-à-dire des tests qui permettent de conclure si la fonction scédastique est correctement spécifiée ou non. Les résultats d'une telle analyse seraient très similaires à ceux du Chapitre 12. En particulier, nous trouverions que les tests dans des directions scédastiques localement équivalentes aux directions dans lesquelles le DGP diffère de l'hypothèse nulle auraient les plus grands paramètres de non centralité (ou NCP) parmi tous ces tests, et que, pour tout NCP, la puissance d'un test serait inversement reliée à son nombre de degrés de liberté. Les lecteurs peuvent trouver là un excellent exercice de démonstration.

Dans cette section, nous nous occuperons d'un problème différent. Qu'advient-il si l'on teste dans une direction scédastique lorsque la fonction scédastique est correctement spécifiée mais que la fonction de régression ne l'est pas? Il semble clair qu'une quelconque mauvaise spécification de la fonction de régression engendrera des résidus "instables". Dans de nombreux cas, si nous oublions un régresseur dont la variance est non constante, par exemple, les résidus seront hétéroscédastiques. Ainsi, il semblerait que les tests de certaines formes d'hétéroscédasticité constitueraient un bon moyen de détecter la mauvaise spécification des fonctions de régression. Il s'avère que ce n'est pas le cas, comme nous allons le voir.

Soit le modèle qui nous intéresse

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}) + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}.$$

Comme dans la Section 12.5, nous supposons que les données sont effectivement générées par une dérive de DGP de la forme

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0) + \alpha n^{-1/2} \mathbf{a} + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma_0^2 \mathbf{I}. \quad (16.35)$$

Ici  $\boldsymbol{\beta}_0$  et  $\sigma_0^2$  désignent les valeurs particulières de  $\boldsymbol{\beta}$  et  $\sigma^2$ ,  $\mathbf{a}$  un vecteur de dimension  $n$  qui peut dépendre de variables exogènes, du vecteur paramétrique  $\boldsymbol{\beta}_0$ , et peut-être même des valeurs passées de  $y_t$ , et  $\alpha$  un paramètre qui détermine l'éloignement du DGP de l'hypothèse nulle

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\beta}_0) + \mathbf{u}, \quad E(\mathbf{u}) = \mathbf{0}, \quad E(\mathbf{u}\mathbf{u}^\top) = \sigma_0^2 \mathbf{I}. \quad (16.36)$$

La dérive de DGP (16.35) tend vers cette hypothèse nulle lorsque  $n \rightarrow \infty$ . Comme nous l'avons vu dans la Section 12.3, le vecteur  $\mathbf{a}$  peut être spécifié de plusieurs manières afin de correspondre au mieux à n'importe quelle sorte de mauvaise spécification de  $\mathbf{x}(\boldsymbol{\beta})$ .

A présent, voyons ce qu'il advient lorsque nous testons l'hypothèse nulle où les  $u_t$  sont homoscedastiques contre l'alternative

$$E(u_t^2) = h(\alpha + \mathbf{Z}_t \boldsymbol{\gamma}),$$

où  $\mathbf{Z}_t$  est un vecteur de dimension  $1 \times q$ . Si nous ne supposons pas que les aléas sont normalement distribués, une statistique de test possible (de la forme  $\chi^2$ ) est  $n$  fois le  $R^2$  centré de la régression de  $\hat{\mathbf{v}}$ , un vecteur d'élément type  $\hat{u}_t^2$ , sur une constante et  $\mathbf{Z}$ . Cette statistique de test peut s'écrire

$$\begin{aligned} & \frac{\hat{\mathbf{v}}^\top \mathbf{M}_\iota \mathbf{Z} (\mathbf{Z}^\top \mathbf{M}_\iota \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{M}_\iota \hat{\mathbf{v}}}{n^{-1} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \hat{\mathbf{v}}} \\ &= \frac{(n^{-1/2} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \mathbf{Z}) (n^{-1} \mathbf{Z}^\top \mathbf{M}_\iota \mathbf{Z})^{-1} (n^{-1/2} \mathbf{Z}^\top \mathbf{M}_\iota \hat{\mathbf{v}})}{n^{-1} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \hat{\mathbf{v}}}, \end{aligned} \quad (16.37)$$

où  $\mathbf{M}_\iota$  désigne la matrice qui calcule les écarts à la moyenne. Pour connaître la distribution asymptotique de (16.37) sous le DGP (16.35), nous devons examiner les quantités

$$n^{-1/2} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \mathbf{Z}, \quad n^{-1} \mathbf{Z}^\top \mathbf{M}_\iota \mathbf{Z}, \quad \text{et} \quad n^{-1} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \hat{\mathbf{v}}$$

lorsque  $n \rightarrow \infty$ . Pour la seconde d'entre elles, nous supposons que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{M}_\iota \mathbf{Z} \right)$$

existe et est une matrice définie positive. La question se ramène donc au comportement des deux autres expressions,  $n^{-1/2} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \mathbf{Z}$  et  $n^{-1} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \hat{\mathbf{v}}$ .

Dans la Section 12.4, nous avons obtenu le résultat que, sous un DGP tel (16.35),

$$\hat{\mathbf{u}} \equiv \mathbf{y} - \hat{\mathbf{x}} = \mathbf{M}_X (\mathbf{u} + \alpha n^{-1/2} \mathbf{a}) + o(n^{-1/2}), \quad (16.38)$$

où  $\mathbf{M}_X$  désigne la matrice qui projette sur  $\mathcal{S}^\perp(\mathbf{X}_0)$ . En exploitant ce résultat, nous pouvons montrer qu'à la fois  $n^{-1/2} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \mathbf{Z}$  et  $n^{-1} \hat{\mathbf{v}}^\top \mathbf{M}_\iota \hat{\mathbf{v}}$  tendent vers les mêmes quantités que  $\alpha$  soit nul ou non. La démonstration est légèrement ennuyeuse mais nullement difficile, et les lecteurs peuvent la trouver intéressante à travailler. Ainsi, nous concluons que la statistique de test (16.37) doit avoir la même distribution asymptotique — à savoir  $\chi^2(q)$  — sous (16.35) comme sous (16.36). Elle aura donc une puissance asymptotique égale à son niveau.

Ce résultat peut sembler à première vue plutôt remarquable. Il indique que si la taille d'échantillon est suffisamment grande, et que le DGP diffère de l'hypothèse nulle d'une quantité proportionnelle à  $n^{-1/2}$ , alors un test dans n'importe quelle direction scédastique aura une puissance égale à son niveau. Par contraste, un test dans n'importe quelle direction de régression non orthogonale à  $\mathbf{M}_X \mathbf{a}$  aura une puissance supérieure à son niveau. Dans la pratique, bien entendu, les tailles d'échantillon ne sont pas infinies et les DGP sont toujours à une distance finie de l'hypothèse nulle, de sorte que nous ne nous attendrions pas à ce que ces résultats soient exacts. Mais ils suggèrent fortement que les tests dans des directions scédastiques seront beaucoup moins puissants que les tests appropriés orientés régression lorsque c'est la fonction de régression qui est mal spécifiée.

**Tableau 16.2** Puissance des Différents Tests lorsque le DGP a des Erreurs AR(1)

$n$	$\rho$	AR(1)	AR(1)	ARCH(1)	ARCH(1)
		Puissance à 1%	5%	1%	5%
50	0.566	66.0	85.3	6.5	14.3
100	0.400	82.2	94.2	8.4	15.8
200	0.283	87.1	96.2	7.1	14.3
400	0.200	89.9	97.1	5.6	11.7
800	0.141	90.8	97.4	3.6	8.7
1600	0.100	90.8	97.5	2.3	7.1
3200	0.071	91.3	97.6	1.8	5.8
6400	0.050	92.0	98.0	1.6	5.7
12800	0.035	92.2	97.9	1.3	5.4

Pour illustrer la façon dont ce résultat asymptotique s'applique en échantillon fini, considérons l'exemple suivant. Les données sont générées par un processus AR(1) avec un terme constant et un paramètre  $\rho$  égal à  $4/n^{1/2}$ . L'hypothèse nulle est que  $y_t$  égale une constante plus des erreurs bruits blancs. Cette hypothèse nulle est testée contre deux hypothèses alternatives à l'aide de tests LM basés sur des régressions artificielles. Ces deux hypothèses alternatives sont que les aléas suivent un processus AR(1) d'une part, ce qui est effectivement le cas, et un processus ARCH(1) d'autre part. Les pourcentages de rejet de l'hypothèse nulle pour les deux tests aux niveaux 1% et 5%, pour différentes tailles d'échantillon et les valeurs correspondantes de  $\rho$  sont rassemblés dans le Tableau 16.2. Ces résultats proviennent d'une expérience Monte Carlo, à 10.000 simulations pour chaque taille d'échantillon.

Nous voyons à partir du Tableau 16.2 que, dans ce cas, le test contre les erreurs AR(1) a toujours une puissance beaucoup plus importante que le test contre les erreurs ARCH(1). Quand la taille de l'échantillon augmente et que  $\rho$  converge vers zéro, la puissance du premier test augmente tout d'abord quelque peu et se stabilise ensuite. Au contraire, la puissance du dernier test augmente légèrement dans un premier temps puis commence à diminuer régulièrement vers sa taille asymptotique à 1% ou 5%. Bien que nous ne les ayons pas discutés, des résultats similaires tiennent pour le cas où la fonction de régression est correctement spécifiée et que la fonction scédastique est mal spécifiée. Si le DGP se rapproche de l'hypothèse nulle à un taux approprié dans ce cas, les tests orientés régression auront asymptotiquement une puissance égale à leur niveau.

Il est important de garder en mémoire les résultats de cette section quand nous testons la spécification d'un modèle de régression. Ils suggèrent fortement que si seule la fonction de régression est mal spécifiée, alors des tests quelconques orientés régression devraient avoir des  $P$ -values beaucoup plus faibles



que des tests dans n'importe quelle direction scédastique. Réciproquement, si seule la fonction scédastique est mal spécifiée, alors des tests dans des directions scédastiques quelconques devraient avoir des  $P$ -values beaucoup plus faibles que des tests dans n'importe quelle direction scédastique. Si, au contraire, les tests dans les directions de régression qui rejettent l'hypothèse nulle le plus souvent ont des  $P$ -values à peu près comparables à celles des tests dans des directions scédastiques qui rejettent l'hypothèse nulle le plus souvent, alors il semble tout à fait probable qu'à la fois la fonction de régression et la fonction scédastique sont toutes deux mal spécifiées.

## 16.7 TESTS D'ASYMÉTRIE ET D'APLATISSEMENT

Bien qu'il soit correct d'utiliser les moindres carrés chaque fois que les aléas associés à une fonction de régression sont d'espérance nulle et ont une matrice de covariance qui satisfait des conditions de régularité assez faibles, les moindres carrés conduisent à un estimateur optimal seulement dans des circonstances particulières. Par exemple, nous avons vu dans le Chapitre 9 que lorsque la matrice de covariance des aléas n'est pas une matrice scalaire, l'estimateur GLS est plus efficace que celui des OLS. Ainsi, l'information sur les seconds moments des aléas conduira, en général, à un gain d'efficacité dans l'estimation des paramètres de la fonction de régression. C'est également le cas pour les moments des aléas d'ordre supérieur. Par exemple, si les aléas sont très leptokurtiques, c'est-à-dire si leur distribution possède des queues très épaisses, les moindres carrés peuvent se révéler très inefficaces par rapport à un autre estimateur qui prend en compte la leptokurtosis. De façon similaire, si les aléas sont asymétriques, il sera possible de trouver mieux que les moindres carrés en utilisant un estimateur qui reconnaît la présence d'asymétrie. Bien sûr, les aléas asymétriques peuvent très bien indiquer que le modèle est mal spécifié; peut-être la variable dépendante devrait-elle être transformée avant l'estimation, par exemple (voir le Chapitre 14).

Tout cela suggère qu'il est généralement prudent de tester l'hypothèse que les aléas sont normalement distribués. Dans la pratique, nous calculons rarement des moments au-delà du troisième ou du quatrième; ceci signifie que l'on teste l'asymétrie ou l'excès de kurtosis (aplatissement). Rappelons d'après la Section 2.6 que, pour une distribution normale de variance  $\sigma^2$ , le troisième moment centré, qui détermine l'asymétrie, est nul, tandis que le quatrième moment centré, qui détermine l'excès de kurtosis, est  $3\sigma^4$ . Si le troisième moment centré n'est pas nul, la distribution est asymétrique. Si le quatrième moment centré est supérieur à  $3\sigma^4$ , la distribution est dite leptokurtique, alors que si le quatrième moment centré est inférieur à  $3\sigma^4$ , la distribution est dite platykurtique. Dans la pratique, les résidus sont fréquemment leptokurtiques et rarement platykurtiques.

Une approche des tests d'hypothèse de normalité consiste à encadrer la distribution normale par une famille de distributions beaucoup plus générale

et ensuite à concevoir des tests LM pour l'hypothèse nulle que les paramètres "encadrés" sont nuls. Lorsque cette approche est employée, comme dans Jarque et Bera (1980) et Kiefer et Salmon (1983), les tests qui en résultent s'avèrent être simplement des tests pour l'asymétrie et l'excès de kurtosis. Aussi ne discuterons-nous pas de cette approche d'"encadrement" mais supposerons dès le début que nous voulons tester les hypothèses

$$E(u_t^3) = 0 \quad \text{et} \quad E(u_t^4) = 3\sigma^4,$$

où  $u_t$  désigne un aléa type, supposé IID(0,  $\sigma^2$ ).

Dans le cas des modèles de régression, les tests pour l'asymétrie et l'excès de kurtosis sont presque toujours basés sur les résidus. Si  $\hat{u}_t$  est le  $t^{\text{ième}}$  résidu d'un modèle de régression avec un terme constant, nous pouvons tester l'asymétrie en cherchant la moyenne et l'écart type empiriques d'un vecteur d'élément type  $\hat{u}_t^3$  et ensuite en construisant un  $t$  de Student (asymptotique) pour l'hypothèse que la véritable espérance est nulle. De façon similaire, nous pouvons tester l'excès de kurtosis en calculant la moyenne et l'écart type empiriques d'un vecteur d'élément type  $\hat{u}_t^4 - 3\hat{\sigma}^4$ , où  $\hat{\sigma}$  est l'estimation ML de  $\sigma$ , et en construisant de la même façon un  $t$  de Student (asymptotique). Cependant, ces procédures peuvent ne pas être les meilleures parce les écarts types estimés utilisés pour construire les statistiques de test ne prennent pas entièrement en compte les implications de l'hypothèse de normalité.

Supposons que les aléas  $u_t$  associés à un quelconque modèle de régression soient distribués selon une NID(0,  $\sigma^2$ ). Alors, si les résidus sont notés  $\hat{u}_t$  et l'estimation ML de la variance est  $\hat{\sigma}^2$ , nous pouvons montrer que (voir ultérieurement)

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \left( \frac{\hat{u}_t^3}{\hat{\sigma}^3} \right)^2 \right) = 6 \quad (16.39)$$

et

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \left( \frac{\hat{u}_t^4}{\hat{\sigma}^4} - 3 \right)^2 \right) = 24. \quad (16.40)$$

Ces deux résultats nous simplifient le calcul des statistiques de test. Nous utilisons simplement les résidus normalisés

$$e_t \equiv \frac{\hat{u}_t - \hat{\mu}}{\hat{\sigma}},$$

où  $\hat{\mu}$  désigne la moyenne empirique des  $\hat{u}_t$  (qui peut être différente de zéro pour les modèles ne comportant pas l'équivalent d'un terme constant). Alors, une statistique de test pour l'asymétrie est

$$(6n)^{-1/2} \sum_{t=1}^n e_t^3 \quad (16.41)$$

et une statistique de test pour l'excès de kurtosis est

$$(24n)^{-1/2} \sum_{t=1}^n (e_t^4 - 3). \quad (16.42)$$

Chacune de ces statistiques de test sera distribuée asymptotiquement selon une  $N(0, 1)$  sous l'hypothèse nulle de normalité. Leurs carrés seront asymptotiquement distribués selon une  $\chi^2(1)$ . De plus, puisque nous pouvons montrer que ces deux statistiques sont indépendantes, la somme de leurs carrés sera asymptotiquement distribuée selon une  $\chi^2(2)$ . Ces deux statistiques de test furent suggérées (sous une forme quelque peu différente) par Jarque et Bera (1980); <sup>4</sup> consulter également White et MacDonald (1980) et Bera et Jarque (1981, 1982).

Nous n'avons pas encore justifié les résultats (16.39) et (16.40). Afin d'y remédier, nous partons d'un résultat standard de la distribution  $N(0, 1)$ . Si une variable aléatoire  $z$  est distribuée selon une  $N(0, 1)$ , alors tous ses moments impairs sont nuls (la distribution est symétrique), et les moments pairs sont donnés par la formule

$$E(z^{2n}) = \prod_{i=1}^n (2i - 1);$$

consulter la Section 4 de l'Annexe B. Une extension facile de ce résultat nous apprend que si  $z$  est distribuée selon une  $N(0, \sigma^2)$ , alors

$$E(z^{2n}) = \sigma^{2n} \prod_{i=1}^n (2i - 1). \quad (16.43)$$

Ainsi, si les résidus normalisés étaient distribués en fait selon une  $NID(0, 1)$ , nous trouverions que le membre de gauche de (16.39) serait égal au sixième moment de la distribution  $N(0, 1)$ , ou 15. De façon similaire, le membre de gauche de (16.40) serait

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n (z^8 - 6z^4 + 9) \right) = 105 - 18 + 9 = 96.$$

L'inexactitude des ces résultats provient du fait que les résidus normalisés sont calculés à l'aide d'*estimations* de l'espérance et de l'écart type.

Pour s'en rendre compte, le moyen le plus simple de procéder est d'imaginer de baser un test sur la régression OPG dont nous avons discuté, pour la

<sup>4</sup> Kiefer et Salmon (1983) ont proposé des statistiques de test qui apparaissent quelque peu différentes mais sont rigoureusement identiques numériquement à (16.41) et (16.42).

première fois dans la Section 13.7. Supposons, pour simplifier, que le modèle de régression à tester soit

$$y_t = \beta + u_t, \quad u_t \sim \text{NID}(0, \sigma^2); \quad (16.44)$$

l'ajout de régresseurs en plus du terme constant ne modifie pas les résultats. Lorsque tous les termes sont évalués avec les véritables valeurs, la régression OPG qui correspond à ce modèle peut s'écrire

$$1 = bu_t + s(u_t^2 - \sigma^2) + \text{résidu}, \quad (16.45)$$

où nous avons remplacé  $y_t - \beta$  par  $u_t$  et multiplié le premier régresseur par  $\sigma^2$  et le second par  $\sigma^3$ . (Puisque nous sommes intéressés par le calcul des statistiques de test, et non par l'estimation des matrices de covariance, il est parfaitement légitime de multiplier n'importe lequel des régresseurs par une constante). Un test de (16.44) contre n'importe quelle alternative peut être basé sur une régression OPG. Nous devons simplement ajouter une ou plusieurs colonnes correctement spécifiées à (16.45).

Pour tester contre l'alternative que les aléas sont asymétriques, le régresseur naturel à ajouter à (16.45) est  $u_t^3$ . La statistique de test sera simplement le  $t$  de Student associé à ce régresseur. Le numérateur de cette statistique est donc simplement la moyenne du régresseur, après qu'il ait été projeté orthogonalement sur le complément orthogonal de l'espace engendré par les deux autres régresseurs. Sous l'hypothèse nulle, le régresseur de test est déjà asymptotiquement orthogonal au second régresseur dans (16.45), puisque tous les moments impairs d'une distribution normale centrée sont nuls. Mais il ne sera pas orthogonal au premier régresseur dans (16.45). La projection de  $u_t^3$  orthogonalement sur le complément orthogonal de l'espace engendré par  $u_t$  donne

$$u_t^3 - u_t \left( \frac{\sum_{t=1}^n u_t^4}{\sum_{t=1}^n u_t^2} \right) \stackrel{a}{=} u_t^3 - 3\sigma^2 u_t. \quad (16.46)$$

Ici, l'égalité asymptotique est obtenue en divisant chaque terme de somme par  $n$  puis en calculant les limites en probabilité. Il est facile de vérifier à partir de (16.43) que la variance de  $u_t^3 - 3\sigma^2 u_t$  est  $6\sigma^6$ .

De façon similaire, pour tester contre l'alternative que les aléas ont un quatrième moment différent de  $3\sigma^4$ , le régresseur naturel à ajouter à (16.45) est  $u_t^4 - 3\sigma^4$ . A nouveau, le numérateur de la statistique de test sera la moyenne de ce régresseur, après qu'il ait été projeté orthogonalement sur le complément orthogonal de l'espace engendré par les deux autres. Dans ce cas, le régresseur de test est asymptotiquement orthogonal au premier régresseur dans (16.45), mais pas au second. La projection orthogonale du régresseur de test sur les derniers donne

$$\begin{aligned} & u_t^4 - 3\sigma^4 - (u_t^2 - \sigma^2) \frac{\sum_{t=1}^n (u_t^6 - u_t^4 \sigma^2 - 3u_t^2 \sigma^4 + 3\sigma^6)}{\sum_{t=1}^n (u_t^4 - 2u_t^2 \sigma^2 + \sigma^4)} \\ & \stackrel{a}{=} u_t^4 - 6u_t^2 \sigma^2 + 3\sigma^4. \end{aligned} \quad (16.47)$$

Ici, l'égalité asymptotique est obtenue de la même façon que dans (16.46). Il est facile de vérifier que la variance de  $u_t^4 - 6u_t^2\sigma^2 + 3\sigma^4$  est  $24\sigma^8$ .

A présent, supposons que nous remplaçons  $u_t$  par  $e_t$  et  $\sigma$  par 1 dans les expressions des membres de droite de (16.46) et (16.47). Alors, il est facile de voir que

$$\sum_{t=1}^n (e_t^3 - 3e_t) = \sum_{t=1}^n e_t^3 \quad \text{et}$$

$$\sum_{t=1}^n (e_t^4 - 6e_t^2 + 3) = \sum_{t=1}^n (e_t^4 - 3).$$

Les membres de droite de ces expressions sont simplement les numérateurs des statistiques de test (16.41) et (16.42). A partir de ces égalités, nous voyons que ces dernières doivent avoir des variances égales à  $n$  fois celles des expressions (16.46) et (16.47) quand  $\sigma = 1$ . Cela explique la provenance des dénominateurs des statistiques de test et complète la démonstration de (16.39) et (16.40).

La démonstration précédente expose clairement pourquoi la prise en compte de régresseurs autres que la constante ne changerait pas le résultat. Si un tel régresseur est noté  $X_t$ , alors la colonne lui correspondant dans la régression OPG (16.45) a pour élément type  $X_t u_t$ . Mais la covariance entre cet élément et  $u_t^3 - 3\sigma^2 u_t$  et  $u_t^4 - 6u_t^2\sigma^2 + 3\sigma^4$  est nulle, de sorte que les colonnes de test sont automatiquement et asymptotiquement orthogonales à n'importe quelle direction de régression. En fait, comme elles sont aussi orthogonales à n'importe quelle direction scédastique, les tests pourraient être utilisés avec des résidus normalisés d'une régression dont la fonction scédastique aurait été estimée.

La raison de la simplicité des tests d'asymétrie et d'excès de kurtosis discutés dans cette section est que, comme nous venons de le voir, nous pouvons largement ignorer le fait que les résidus des modèles de régression dépendent des paramètres estimés. Avec des modèles beaucoup plus généraux, nous ne pouvons pas toujours ignorer ce fait. Il semble possible que des variantes de la régression OPG nous fournissent un moyen commode de tester l'asymétrie et l'excès de kurtosis dans de tels modèles. Comme nous le verrons dans la prochaine section, cela est en effet le cas. De tels tests sont effectivement des cas particuliers d'une classe importante et très générale de tests appelés tests de moments conditionnels.

## 16.8 TESTS DE MOMENTS CONDITIONNELS

Une approche importante pour tester la spécification d'un modèle dont nous n'avons pas encore discuté est de baser directement des tests sur certaines conditions que les aléas d'un modèle satisferaient. De tels tests sont parfois

appelés **tests de spécification du moment** mais sont plus fréquemment appelés **tests de moments conditionnels**, ou **tests CM**. Ils furent tout d'abord suggérés par Newey (1985a) et Tauchen (1985) et développés plus tard par White (1987), Pagan et Vella (1989), Wooldridge (1991a, 1991b), et d'autres. L'idée fondamentale est que si un modèle est correctement spécifié, de nombreuses quantités aléatoires fonctions des aléas devraient avoir des espérances nulles. La spécification d'un modèle permet parfois une conclusion plus forte, selon laquelle de telles fonctions des aléas ont des espérances nulle *conditionnellement* à un ensemble d'information — d'où la terminologie des tests de moments conditionnels.

Puisqu'une espérance est souvent assimilée à un moment, la condition qu'une quantité aléatoire soit d'espérance nulle est généralement appelée **condition du moment**. Même si le moment d'une population est nul, sa contrepartie empirique, que nous appellerons un **moment empirique**, ne le sera (presque) jamais exactement, mais elle ne devrait pas être significativement différente de zéro. Les tests de moments conditionnels sont basés directement sur cette propriété.

Les tests de moments conditionnels peuvent être utilisés pour tester de nombreux aspects différents de la spécification des modèles économétriques. Supposons que la théorie économique ou statistique sous-jacente à un modèle paramétrisé donné indique que pour toute observation  $t$  il existe une certaine fonction de la variable dépendante  $y_t$  et des paramètres du modèle  $\theta$ , disons  $m_t(y_t, \theta)$ , dont l'espérance est nulle lorsque le DGP utilisé pour calculer l'espérance est caractérisé par  $\theta$ . Ainsi, pour tout  $t$  et tout  $\theta$ ,

$$E_{\theta}(m_t(y_t, \theta)) = 0. \quad (16.48)$$

Nous pouvons penser de (16.48) qu'elle exprime une condition sur le moment. En général, les fonctions  $m_t$  peuvent aussi dépendre de variables exogènes ou prédéterminées.

Alors même qu'il existe une fonction différente pour chaque observation, il semble raisonnable, par analogie avec les moments empiriques, d'utiliser l'expression suivante comme contrepartie empirique du moment dans la condition du moment (16.48):

$$m(\mathbf{y}, \hat{\theta}) \equiv \frac{1}{n} \sum_{t=1}^n m_t(y_t, \hat{\theta}), \quad (16.49)$$

où  $\hat{\theta}$  désigne un vecteur d'estimations de  $\theta$ . Ainsi, l'expression (16.49) est une forme du moment empirique. Un test CM à un degré de liberté serait calculé en divisant le moment empirique par une estimation de son écart type et serait asymptotiquement distribué selon la  $N(0, 1)$  sous des conditions de régularité appropriées. Il peut bien sûr y avoir plus d'un moment conditionnel, auquel cas la statistique de test serait calculée comme une forme quadratique des

moments empiriques et de l'estimation de leur matrice de covariance, et cette forme quadratique aurait une distribution asymptotique khi-deux.

Il est clair que les tests d'asymétrie et d'excès de kurtosis dont nous avons discuté dans la section précédente sont des cas particuliers des tests CM. Une condition comme  $E(u_t^3) = 0$  est une condition sur le moment (non conditionnel), et une statistique de test comme (16.41) est simplement la contrepartie empirique du moment concerné, divisée par une estimation de son écart type. Ce qui peut être moins clair est qu'une fois autorisée la possibilité de moments conditionnels, pratiquement tous les tests de spécification traités jusqu'à présent peuvent être conçus comme des tests CM. Par exemple, considérons un modèle de régression linéaire comme

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad (16.50)$$

où  $\mathbf{X}_t\boldsymbol{\beta}$  est spécifiée comme la moyenne de  $y_t$  conditionnellement à un ensemble d'information  $\Omega_t$  quelconque. Si ce modèle est spécifié correctement, l'espérance conditionnelle  $E(u_t | \Omega_t)$  devrait être nulle. Cette condition sur le moment conditionnel implique que  $u_t$  devrait être orthogonal à n'importe quelle variable appartenant à  $\Omega_t$ . D'où, pour tout  $z_t \in \Omega_t$ , le moment non conditionnel  $E(u_t z_t)$  devrait être nul. Le moment empirique correspondant est

$$\sum_{t=1}^n \hat{u}_t z_t = \hat{\mathbf{u}}^\top \mathbf{z}, \quad (16.51)$$

où  $\hat{u}_t$  désigne l'estimation de  $u_t$  issue d'une estimation OLS de (16.50),  $\hat{\mathbf{u}}$  est un vecteur de dimension  $n$  d'élément type  $\hat{u}_t$ , et  $\mathbf{z}$  est un vecteur de dimension  $n$  d'élément type  $z_t$ . Nous ne nous sommes pas embarrassés à diviser l'équation (16.51) par  $n$ , puisqu'elle doit être divisée par une quantité qui estime son écart type de manière convergente afin d'obtenir une statistique de test du moment conditionnel.

L'expression (16.51) est, bien sûr, le numérateur d'un  $t$  de Student ordinaire pour  $\gamma = 0$  dans la régression

$$y_t = \mathbf{X}_t\boldsymbol{\beta} + \gamma z_t + u_t. \quad (16.52)$$

Le dénominateur de ce  $t$  de Student est un estimateur convergent de l'écart type de (16.51). Donc le  $t$  de Student ordinaire pour  $\gamma = 0$  dans (16.52) peut être assimilé à un test CM, comme c'est le cas pour une variété de statistiques de test plus compliquées qui estiment la variance de (16.51) selon différentes méthodes. Par exemple, nous pourrions utiliser un des HCCME traités dans la Section 16.3 pour obtenir une statistique de test valable en présence d'hétéroscédasticité de forme inconnue. Cela pourrait être soit un test de type Wald, pour lequel les estimations de la matrice de covariance seraient basées sur des estimations OLS du modèle non contraint (16.52), soit un test de type LM, pour lequel elles seraient basées sur des estimations OLS d'un modèle contraint (16.50), comme dans le cas de la HRGMR.

Ces exemples suggèrent que les tests CM obtenus en explicitant les conditions du moment sont fréquemment ceux que nous connaissons déjà à travers d'autres approches. C'est en effet souvent le cas. Par exemple, considérons le test de l'hypothèse que le vecteur  $\theta_2$  de dimension  $r$  soit nul dans un modèle estimé par maximum de vraisemblance. Les "conditions du moment" naturelles à utiliser sont

$$E(g_i(\theta_1, \mathbf{0})) = 0 \quad \text{pour } i = k - r + 1, \dots, k.$$

Ces conditions établissent que les  $r$  éléments du vecteur score  $\mathbf{g}(\theta)$  correspondant aux éléments de  $\theta_2$  doivent être d'espérance nulle sous l'hypothèse nulle où  $\theta_2 = 0$ . Pour obtenir des moments empiriques, nous remplaçons simplement  $\theta_1$  par les estimations ML contraintes  $\tilde{\theta}_1$ . Cela fournit le vecteur score  $\mathbf{g}(\tilde{\theta})$ . Par conséquent, dans ce cas, la forme score familière d'un test LM peut être assimilée à un test CM.

Dans ces cas et dans de nombreux autres, les tests CM s'avèrent souvent être exactement les mêmes que des tests de spécification plus familiers basés sur les principes LM ou DWH. Quels sont alors les avantages liés à des tests CM par rapport aux autres tests? Pagan et Vella (1989) avancent l'idée, initialement dans le contexte des modèles à variable dépendante limitée, qu'il est souvent beaucoup plus facile et naturel d'écrire des conditions du moment plausibles plutôt que de dériver des tests LM. En effet, c'est fréquemment le cas. Nous avons implicitement suivi l'approche CM dans la section précédente, lorsque nous avons dérivé des tests de normalité en testant explicitement l'asymétrie et l'excès de kurtosis, plutôt que de formuler un modèle alternatif et d'en dériver des tests LM. Comme nous l'avons remarqué, nous aurions pu obtenir des statistiques de tests identiques en utilisant la dernière approche, mais cela nous aurait réclamé plus de travail. Ainsi, l'approche CM peut être attrayante lorsqu'il est facile d'écrire les conditions du moment que l'on souhaite tester ainsi que leurs contreparties empiriques.

Cependant, le simple fait d'écrire un ensemble de moments empiriques ne nous permet pas, à lui seul, d'obtenir une statistique de test. Nous devons aussi pouvoir estimer leur matrice de covariance. Comme nous le verrons de façon très brève, si nous manipulons un modèle estimé par maximum de vraisemblance, auquel s'applique la régression OPG familière, il est possible d'obtenir cette matrice de façon mécanique en utilisant cette régression. Mais bien que cette procédure nous permette d'obtenir des tests CM directement à partir de la régression OPG, ces tests, comme d'autres calculés à partir de régressions OPG, ont fréquemment d'assez mauvaises propriétés en échantillon fini. Si l'on espère obtenir des tests CM ayant de bonnes propriétés en échantillon fini, on peut être obligé d'entreprendre une analyse détaillée, comme celle menée dans la dernière section, sur la distribution des moments empiriques. De façon alternative, dans certains cas, les tests peuvent être calculés à l'aide de régressions artificielles ayant de meilleures propriétés en échantillon fini que celles de la régression OPG. La conclusion de tout ceci est que l'obtention des tests CM n'est pas toujours une chose facile à gérer.



A présent, nous discutons d'un résultat important, de Newey (1985a), qui permet de calculer des tests CM à l'aide d'une régression OPG. Supposons, pour simplifier, que nous soyons intéressés par le test d'une seule condition du moment, disons  $E_{\theta}(m_t(y_t, \theta)) = 0$ . Le moment empirique correspondant est  $m(\mathbf{y}, \hat{\theta})$ , défini dans l'expression (16.49). Si nous connaissions la véritable valeur de  $\theta$ , il serait clairement très facile d'obtenir une statistique de test. Nous aurions simplement besoin que la condition CLT (voir la Définition 4.16) s'applique à  $n^{1/2}m(\mathbf{y}, \theta)$ , et pourrions ensuite estimer la variance asymptotique de cette expression par

$$\frac{1}{n} \sum_{t=1}^n m_t^2(y_t, \theta). \quad (16.53)$$

La construction d'une statistique de test asymptotiquement distribuée selon la  $N(0, 1)$  serait alors très facile. Le problème est que dans la plupart des cas, nous ne connaissons pas  $\theta$  mais seulement un vecteur d'estimations ML  $\hat{\theta}$ . Comme nous le verrons dans un instant, la variance asymptotique de  $n^{1/2}m(\mathbf{y}, \hat{\theta})$  est généralement plus petite que  $n^{1/2}m(\mathbf{y}, \theta)$ , de sorte qu'il n'est généralement pas correct d'estimer la première en utilisant simplement  $\hat{\theta}$  à la place de  $\theta$  dans (16.53).

Nous commençons par calculer un développement de Taylor au premier ordre de  $n^{1/2}m(\mathbf{y}, \theta)$  par rapport au vecteur  $\theta$  de dimension  $k$  autour du véritable vecteur paramétrique  $\theta_0$ . Le résultat, lorsque nous l'évaluons en  $\theta = \hat{\theta}$ , est

$$n^{1/2}m(\mathbf{y}, \hat{\theta}) \cong n^{1/2}m(\mathbf{y}, \theta_0) + \mu_0^\top n^{1/2}(\hat{\theta} - \theta_0). \quad (16.54)$$

Ici  $\mu_0$  désigne le vecteur des dérivées partielles de  $m(\mathbf{y}, \theta)$  par rapport à  $\theta$ , évalué en  $\theta_0$ . Comme chacun des termes dans (16.54) est  $O(1)$ , les différences entre  $m(\mathbf{y}, \hat{\theta})$  et  $m(\mathbf{y}, \theta_0)$  ne peuvent pas être ignorées asymptotiquement.

Ensuite, nous obtenons un résultat général et utile qui nous permettra de remplacer le vecteur des dérivées  $\mu_0$  dans (16.48) par une grandeur plus commode. La condition sur le moment testée est fournie par (16.48). Le moment peut explicitement s'écrire comme

$$E_{\theta}(m_t(y_t, \theta)) = \int_{-\infty}^{\infty} m_t(y_t, \theta) L_t(y_t, \theta) dy_t. \quad (16.55)$$

En différentiant le membre de droite de (16.55) par rapport aux composantes de  $\theta$ , nous obtenons, en raisonnant de la même façon que pour l'égalité (8.44) de la matrice d'information,

$$E_{\theta}(m_t(\theta) \mathbf{G}_t(\theta)) = -E_{\theta}(\mathbf{N}_t(\theta)). \quad (16.56)$$

Ici  $\mathbf{G}_t(\theta)$  est la contribution de l'observation  $t$  au gradient de la fonction de logvraisemblance, et le vecteur ligne  $\mathbf{N}_t(\theta)$  de dimension  $1 \times k$  a pour élément type  $\partial m_t(\theta) / \partial \theta_i$ .<sup>5</sup> La forme la plus intéressante de notre résultat est obtenue

<sup>5</sup> Notre notation usuelle aurait dû être  $\mathbf{M}_t(\theta)$  à la place de  $\mathbf{N}_t(\theta)$ , mais celui-là aurait nui à la distinction d'avec la notation standard des projections orthogonales complémentaires.

en sommant (16.56) sur  $t$ . Soit  $\mathbf{m}(\boldsymbol{\theta})$  un vecteur de dimension  $n$  d'élément type  $m_t(\boldsymbol{\theta})$ , et soit  $\mathbf{N}(\boldsymbol{\theta})$  une matrice de dimension  $n \times k$  de ligne type  $\mathbf{N}_t(\boldsymbol{\theta})$ . Alors

$$\frac{1}{n}E_{\boldsymbol{\theta}}(\mathbf{G}^{\top}(\boldsymbol{\theta})\mathbf{m}(\boldsymbol{\theta})) = -\frac{1}{n}E_{\boldsymbol{\theta}}(\mathbf{N}^{\top}(\boldsymbol{\theta})\boldsymbol{\iota}), \quad (16.57)$$

où, comme d'habitude,  $\mathbf{G}(\boldsymbol{\theta})$  désigne la matrice CG. Dans (16.54),  $\boldsymbol{\mu}_0 = n^{-1}\mathbf{N}_0^{\top}\boldsymbol{\iota}$ , où  $\mathbf{N}_0 \equiv \mathbf{N}(\boldsymbol{\theta}_0)$ . D'après la loi des grands nombres,  $\boldsymbol{\mu}_0$  convergera vers la limite du membre de droite de (16.57), et donc aussi vers la limite du membre de gauche. Ainsi, si  $\mathbf{G}_0 \equiv \mathbf{G}(\boldsymbol{\theta}_0)$ , nous pouvons affirmer que

$$\boldsymbol{\mu}_0 = \frac{1}{n}\mathbf{N}_0^{\top}\boldsymbol{\iota} \stackrel{a}{=} -\frac{1}{n}\mathbf{G}_0^{\top}\mathbf{m}_0. \quad (16.58)$$

Ensuite, nous exploitons le résultat très bien connu (13.18) sur la relation liant les estimations ML, la matrice d'information et le vecteur score:

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} \mathcal{I}_0^{-1}n^{-1/2}\mathbf{g}_0. \quad (16.59)$$

Puisque la matrice d'information  $\mathcal{I}_0$  est asymptotiquement égale à  $n^{-1}\mathbf{G}_0^{\top}\mathbf{G}_0$  (voir la Section 8.6), et  $\mathbf{g}_0 = \mathbf{G}_0^{\top}\boldsymbol{\iota}$ , (16.59) devient

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} (n^{-1}\mathbf{G}_0^{\top}\mathbf{G}_0)^{-1}n^{-1/2}\mathbf{G}_0^{\top}\boldsymbol{\iota}.$$

Ce résultat combiné à (16.58) nous permet de remplacer le membre de droite de (16.54) par

$$n^{-1/2}\mathbf{m}_0^{\top}\boldsymbol{\iota} - n^{-1}\mathbf{m}_0^{\top}\mathbf{G}_0(n^{-1}\mathbf{G}_0^{\top}\mathbf{G}_0)^{-1}n^{-1/2}\mathbf{G}_0^{\top}\boldsymbol{\iota} = n^{-1/2}\mathbf{m}_0^{\top}\mathbf{M}_G\boldsymbol{\iota}, \quad (16.60)$$

où  $\mathbf{M}_G$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^{\perp}(\mathbf{G}_0)$ .

Le résultat (16.60) fait apparaître clairement la différence entre le moment empirique évalué avec le  $\boldsymbol{\theta}_0$  inconnu et le même moment évalué avec l'estimation ML  $\hat{\boldsymbol{\theta}}$ , c'est-à-dire la différence entre  $n^{-1/2}\mathbf{m}_0^{\top}\boldsymbol{\iota}$  et  $n^{-1/2}\hat{\mathbf{m}}^{\top}\boldsymbol{\iota}$ . L'effet d'utiliser des estimations est une projection orthogonale implicite du vecteur  $\mathbf{m}_0$  sur le complément orthogonal de l'espace  $\mathcal{S}(\mathbf{G}_0)$  associé aux paramètres du modèle. C'est cette projection qui réduit la variance de l'expression que nous pouvons effectivement calculer par rapport à la variance de l'expression correspondante basée sur les véritables paramètres. Les variances utilisées dans les tests d'asymétrie et d'excès de kurtosis considérés dans la dernière section peuvent aussi être calculées en utilisant (16.60).

A présent, nous sommes prêts à obtenir une expression appropriée de la variance asymptotique de  $n^{-1/2}\hat{\mathbf{m}}^{\top}\boldsymbol{\iota}$ . Nous avons besoin, comme nous l'avons suggéré auparavant, que  $n^{-1/2}\mathbf{m}_0^{\top}\boldsymbol{\iota}$  satisfasse la condition CLT et que, au voisinage de  $\boldsymbol{\theta}_0$ ,  $n^{-1}\mathbf{m}^{\top}(\boldsymbol{\theta})\mathbf{G}_i(\boldsymbol{\theta})$  satisfasse la condition WULLN (Définition 4.17) pour tout  $i = 1, \dots, k$ . La variance asymptotique est alors

clairement  $\text{plim}(n^{-1}\mathbf{m}_0^\top \mathbf{M}_G \mathbf{m}_0)$ , qui peut être estimée de manière convergente par  $n^{-1}\hat{\mathbf{m}}^\top \hat{\mathbf{M}}_G \hat{\mathbf{m}}$ . Cela suggère l'usage de la statistique de test

$$\frac{n^{-1/2}\hat{\mathbf{m}}^\top \boldsymbol{\iota}}{(n^{-1}\hat{\mathbf{m}}^\top \hat{\mathbf{M}}_G \hat{\mathbf{m}})^{1/2}} = \frac{\hat{\mathbf{m}}^\top \boldsymbol{\iota}}{(\hat{\mathbf{m}}^\top \hat{\mathbf{M}}_G \hat{\mathbf{m}})^{1/2}}, \quad (16.61)$$

qui sera asymptotiquement distribuée selon la  $N(0, 1)$ .

La connexion avec la régression OPG est à présent évidente. La statistique de test (16.61) est *presque* le  $t$  de Student associé au coefficient  $b$  de la régression OPG suivante:

$$\boldsymbol{\iota} = \hat{\mathbf{G}}\mathbf{c} + b\hat{\mathbf{m}} + \text{résidus}. \quad (16.62)$$

Asymptotiquement, la statistique (16.61) et le  $t$  de Student de (16.62) sont équivalents, parce la somme des carrés des résidus de (16.62) tend vers  $n$  pour de grandes tailles d'échantillon sous l'hypothèse nulle: les régresseurs  $\hat{\mathbf{G}}$  sont toujours orthogonaux à  $\boldsymbol{\iota}$ , et  $\hat{\mathbf{m}}$  est orthogonal à  $\boldsymbol{\iota}$  si la condition sur le moment est satisfaite. Ce résultat est très encourageant. Sans le régresseur  $\hat{\mathbf{m}}$ , qui est le vecteur servant à définir le moment empirique, la régression (16.62) serait simplement la régression OPG associée au modèle originel, et la SSR serait toujours égale à  $n$ . Ainsi la version OPG du test CM, comme tous les autres tests traités et exécutés à partir de régressions artificielles, est simplement un test de la pertinence des coefficients associés à un ou plusieurs régresseurs de test.

Nous savons maintenant clairement comment étendre les tests CM à un ensemble composé d'une ou plusieurs conditions sur les moments. Nous créons simplement un régresseur de test pour chacun des moments empiriques pour construire une matrice  $\hat{\mathbf{R}} \equiv \mathbf{R}(\hat{\boldsymbol{\theta}})$  de dimension  $n \times r$ , où  $r$  est le nombre de conditions sur les moments. Nous utilisons alors la somme des carrés expliqués de la régression OPG

$$\boldsymbol{\iota} = \hat{\mathbf{G}}\mathbf{c} + \hat{\mathbf{R}}\mathbf{b} + \text{résidus}$$

ou n'importe quel autre test de l'hypothèse artificielle  $\mathbf{b} = \mathbf{0}$  asymptotiquement équivalent. A présent, il est clair comme nous l'avons suggéré auparavant, qu'un test exécuté à l'aide d'une régression OPG peut être interprété comme un test CM. Nous devons simplement interpréter les colonnes de test dans la régression comme des moments empiriques.

Une variante intéressante de la régression de test a été suggérée par Tauchen (1985). En effet, il a permuté la régressande  $\boldsymbol{\iota}$  et le régresseur de test  $\hat{\mathbf{m}}$  afin d'obtenir la régression

$$\hat{\mathbf{m}} = \hat{\mathbf{G}}\mathbf{c}^* + b^*\boldsymbol{\iota} + \text{résidus}. \quad (16.63)$$

La statistique de test est le  $t$  de Student pour  $b^* = 0$ . Elle est numériquement identique au  $t$  de Student sur  $b$  dans (16.62). Ce fait provient du résultat que

nous avons obtenu dans la Section 12.7, auquel nous apportons à présent une démonstration géométrique. Appliquons le Théorème FWL à la fois à (16.62) et (16.63) pour obtenir les deux régressions

$$(16.64) \quad \begin{aligned} \hat{M}_G \boldsymbol{\iota} &= b(\hat{M}_G \hat{\mathbf{m}}) + \text{résidus} \quad \text{et} \\ \hat{M}_G \hat{\mathbf{m}} &= b^*(\hat{M}_G \boldsymbol{\iota}) + \text{résidus}. \end{aligned}$$

Ce sont deux régressions univariées à  $n$  observations. Le  $t$  de Student simple de chacune de ces régressions est obtenu par le produit du même facteur scalaire,  $(n-1)^{1/2}$ , et de la cotangente de l'angle entre la régressande et le régresseur (consulter l'Annexe A). Puisque l'angle est inchangé quand le régresseur et la régressande sont permutés, le  $t$  de Student n'est pas modifié non plus. Le Théorème FWL implique que les  $t$  de Student des première et seconde lignes de (16.64) soient égaux à ceux de la régression OPG (16.62) et de la régression de Tauchen (16.63), respectivement, fois la même correction par les degrés de liberté. Ainsi, nous concluons que les  $t$  de Student basés sur les deux dernières régressions sont numériquement identiques.

Puisque les conditions du premier ordre pour  $\hat{\boldsymbol{\theta}}$  impliquent que  $\boldsymbol{\iota}$  est orthogonal à toutes les colonnes de  $\hat{\mathbf{G}}$ , l'estimation OLS de  $b^*$  dans (16.63) sera égale à la moyenne empirique des éléments de  $\hat{\mathbf{m}}$ . Il en serait ainsi si les régresseurs  $\hat{\mathbf{G}}$  étaient supprimés de la régression. Cependant, comme  $\boldsymbol{\theta}$  a été estimé, ces régresseurs doivent être compris si nous voulons obtenir une estimation correcte de la variance de la moyenne d'échantillon. Comme c'est le cas avec d'autres régressions artificielles étudiées, l'omission des régresseurs qui correspondent aux paramètres estimés sous l'hypothèse nulle résulte en une statistique de test trop petite asymptotiquement.

Réitérons nos tout premiers avertissements à propos de la régression OPG. Comme nous l'avons déjà souligné lorsque nous les avons introduites dans la Section 13.7, les statistiques de test basées sur la régression OPG ont souvent de mauvaises propriétés en échantillon fini. Elles tendent à rejeter l'hypothèse nulle trop souvent lorsqu'elle est vraie. Cela est valable pour les tests CM comme pour les tests LM ou les tests  $C(\alpha)$ . Nous devrions utiliser, si possible, des tests alternatifs ayant de bien meilleures propriétés en échantillon fini, tels les tests basés sur la GNR, la HRGNR, la DLR (Section 14.4), ou la BRMR (Section 15.4), lorsque ces procédures sont applicables. Bien sûr, elles le seront en général seulement si le test CM peut être reformulé comme un test ordinaire, avec une hypothèse alternative explicite à partir de laquelle les régresseurs de test peuvent être générés. S'il n'est pas possible de reformuler le test CM de cette manière, et que l'on doit utiliser la régression OPG, il faudra être très prudent lorsqu'un test suggère de rejeter l'hypothèse nulle. Il est souvent plus sage de vérifier les propriétés en échantillon fini à l'aide des expériences Monte Carlo (voir le Chapitre 21).

## 16.9 TESTS DE LA MATRICE D'INFORMATION

Un type important de test de moments conditionnels est la classe des tests appelés tests de la **matrice d'information**, ou **tests IM**. Ils ont été suggérés en premier par White (1982), bien que l'interprétation de moment conditionnel soit assez récente; voir Newey (1985a) et White (1987). L'idée fondamentale est très simple. Si un modèle estimé par maximum de vraisemblance est correctement spécifié, la matrice d'information doit être asymptotiquement égale à l'opposée de la matrice Hessienne. Si le modèle n'est pas correctement spécifié, cette égalité ne sera pas vraie en général, parce que la démonstration de l'égalité de la matrice d'information dépend crucialement du fait que la densité jointe des données est la fonction de vraisemblance; consulter la Section 8.6.

Considérons un modèle statistique caractérisé par une fonction de logvraisemblance de la forme

$$\ell(\mathbf{y}, \boldsymbol{\theta}) = \sum_{t=1}^n \ell_t(y_t, \boldsymbol{\theta}),$$

où  $\mathbf{y}$  désigne un vecteur de  $n$  observations  $y_t$ ,  $t = 1, \dots, n$ , sur une variable dépendante, et  $\boldsymbol{\theta}$  désigne un vecteur de  $k$  paramètres. Comme l'indique l'indice  $t$ , la contribution  $\ell_t$  de l'observation  $t$  à la fonction de logvraisemblance peut dépendre de variables exogènes ou prédéterminées qui varient entre les  $n$  observations. L'hypothèse nulle pour le test IM est que

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \left( \frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_j} \right) \right) = 0, \quad (16.65)$$

pour  $i = 1, \dots, k$  et  $j = 1, \dots, i$ . L'expression (16.65) est un élément type de l'égalité de la matrice d'information. Le premier terme est un élément de la matrice Hessienne, et le second est l'élément correspondant du produit extérieur du gradient. Puisque le nombre de tels termes est  $\frac{1}{2}k(k+1)$ , le nombre de degrés de liberté pour un test IM est potentiellement très grand.

Sans la limite en probabilité, le membre de gauche de (16.65) ressemble à un moment empirique. Cela suggère, concrètement, que l'on peut calculer des tests IM à l'aide d'une régression OPG, procédure suggérée initialement par Chesher (1983) et Lancaster (1984). Nous devons simplement construire une matrice  $\mathbf{Z}(\boldsymbol{\theta})$  de dimension  $n \times \frac{1}{2}k(k+1)$  d'élément type

$$\frac{\partial^2 \ell_t(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} + \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ell_t(\boldsymbol{\theta})}{\partial \theta_j}$$

et l'évaluer avec les estimations ML  $\hat{\boldsymbol{\theta}}$  pour obtenir  $\hat{\mathbf{Z}}$ . Ensuite, nous calculons une régression OPG, avec les régresseurs  $\hat{\mathbf{G}}$  et  $\hat{\mathbf{Z}}$ , puis utilisons  $n$  moins la SSR comme statistique de test. A condition que la matrice  $[\hat{\mathbf{G}} \ \hat{\mathbf{Z}}]^\top [\hat{\mathbf{G}} \ \hat{\mathbf{Z}}]$

soit de plein rang asymptotiquement, la statistique de test sera distribuée asymptotiquement selon une  $\chi^2(\frac{1}{2}k(k+1))$ . Lorsque des colonnes quelconques de  $\hat{\mathbf{G}}$  et  $\hat{\mathbf{Z}}$  sont parfaitement colinéaires, comme cela est souvent le cas, le nombre de degrés de liberté pour le test doit être réduit, bien entendu, en conséquence.

Il est intéressant de considérer comme exemple le modèle de régression non linéaire univarié

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad u_t \sim \text{NID}(0, \sigma^2),$$

où  $x_t(\boldsymbol{\beta})$  est une fonction deux fois continuellement différentiable qui dépend de  $\boldsymbol{\beta}$ , un vecteur de  $p$  paramètres, et de variables exogènes et prédéterminées qui varient suivant les observations. Ainsi, le nombre total de paramètres est  $k = p + 1$ . Pour ce modèle, la contribution à la fonction de logvraisemblance de la  $t^{\text{ième}}$  observation est

$$\ell_t(\boldsymbol{\beta}, \sigma) = -\frac{1}{2} \log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2} (y_t - x_t(\boldsymbol{\beta}))^2.$$

Ainsi, la contribution de la  $t^{\text{ième}}$  observation du régresseur correspondant au  $i^{\text{ième}}$  élément de  $\boldsymbol{\beta}$  est

$$G_{ti}(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^2} (y_t - x_t(\boldsymbol{\beta})) X_{ti}(\boldsymbol{\beta}), \quad (16.66)$$

où, comme d'habitude,  $X_{ti}(\boldsymbol{\beta})$  désigne la dérivée partielle de  $x_t(\boldsymbol{\beta})$  par rapport à  $\beta_i$ . De façon similaire, la contribution de la  $t^{\text{ième}}$  observation au régresseur correspondant à  $\sigma$  est

$$G_{t,k}(\boldsymbol{\beta}, \sigma) = -\frac{1}{\sigma} + \frac{1}{\sigma^3} (y_t - x_t(\boldsymbol{\beta}))^2. \quad (16.67)$$

En utilisant (16.66) et (16.67), il est facile de calculer les régresseurs pour la version OPG du test IM. Nous avons les définitions

$$\hat{e}_t \equiv \frac{1}{\hat{\sigma}} (y_t - x_t(\hat{\boldsymbol{\beta}})), \quad \hat{X}_{ti} \equiv X_{ti}(\hat{\boldsymbol{\beta}}), \quad \text{et} \quad X_{tij}^*(\boldsymbol{\beta}) \equiv \frac{\partial X_{ti}(\boldsymbol{\beta})}{\partial \beta_j}.$$

Alors, à des facteurs multiplicatifs près qui peuvent être sans effet sur l'ajustement de la régression, et donc sans effet sur la valeur de la statistique de test IM, les régresseurs pour la régression de test sont

$$\text{pour } \beta_i : \quad \hat{e}_t \hat{X}_{ti}; \quad (16.68)$$

$$\text{pour } \sigma : \quad \hat{e}_t^2 - 1; \quad (16.69)$$

$$\text{pour } \beta_i \times \beta_j : \quad (\hat{e}_t^2 - 1) \hat{X}_{ti} \hat{X}_{tj} + \hat{\sigma} \hat{e}_t \hat{X}_{tij}^*; \quad (16.70)$$

$$\text{pour } \sigma \times \beta_i : \quad (\hat{e}_t^3 - 3\hat{e}_t) \hat{X}_{ti}; \quad (16.71)$$

$$\text{pour } \sigma \times \sigma : \quad \hat{e}_t^4 - 5\hat{e}_t^2 + 2. \quad (16.72)$$

Les expressions (16.68) et (16.69) définissent les éléments de chaque ligne de  $\hat{\mathbf{G}}$ , alors que les expressions (16.70)–(16.72) définissent les éléments de chaque ligne de  $\hat{\mathbf{Z}}$ . Lorsque la régression originelle comprend un terme constant, (16.69) sera parfaitement colinéaire à (16.70) quand  $i$  et  $j$  indicent la constante. Donc, cette dernière doit être supprimée et le degré de liberté pour le test réduit de 1 à  $\frac{1}{2}(p+2)(p+1) - 1$ .

Les expressions (16.68)–(16.72) révèlent les formes de mauvaise spécification que le test IM teste dans le contexte de la régression non linéaire. Il est évident à partir de (16.71) que les régresseurs  $(\beta_i, \sigma)$  sont ceux qui correspondent à une asymétrie reliée aux  $\hat{X}_{ti}$ . Il apparaît qu'une telle asymétrie, si elle existe, biaise les estimations des covariances de  $\hat{\beta}$  et  $\hat{\sigma}$ . Si nous ajoutons deux fois (16.69) à (16.72), le résultat est  $\hat{e}_t^4 - 3$ , à partir duquel nous voyons que la part linéairement indépendante du régresseur  $(\sigma, \sigma)$  teste dans la direction de l'excès de kurtosis. Aussi bien la platykurtosis que la leptokurtosis conduiraient à un biais dans l'estimation de la variance de  $\hat{\sigma}$ . Il est évident à partir de (16.70) que si  $x_t(\beta)$  était linéaire, les régresseurs  $(\beta_i, \beta_j)$  testeraient le même type d'hétéroscédasticité que le test de White(1980) est conçu de détecter; voir la Section 16.5. Cependant, dans le cas d'une régression non linéaire considéré ici, ces régresseurs testent en même temps la mauvaise spécification d'une fonction de régression. Pour davantage de détails sur ce cas particulier des modèles de régression linéaire, voir Hall (1987).

L'analyse précédente suggère que, dans le cas des modèles de régression, il est probablement plus attrayant de tester directement l'hétéroscédasticité, l'asymétrie, l'excès de kurtosis, et la mauvaise spécification de la fonction de régression plutôt que d'utiliser un test IM. Nous avons déjà vu la façon de tester chacun de ces types de mauvaise spécification individuellement. Les tests individuels peuvent être bien plus puissants et plus informatifs qu'un test IM, en particulier si seuls quelques éléments sont faux dans le modèle. Si on s'intéresse principalement aux inférences sur  $\beta$ , alors tester l'asymétrie et l'excès de kurtosis peut être facultatif.

Il existe un problème très sérieux avec les tests IM basés sur la régression OPG. En échantillon fini, ils tendent à rejeter l'hypothèse nulle beaucoup trop souvent quand elle est vraie. Dans cette optique, les tests IM semblent même plus mauvais que d'autres tests de spécification basés sur la régression OPG. Les résultats Monte Carlo démontrant la mauvaise performance en échantillon fini de la version OPG du test IM peuvent se trouver chez Taylor (1987), Kennan et Neumann (1988), Orme (1990a), Hall (1990), Chesher et Spady (1991), et Davidson et MacKinnon (1992). Dans certains de ces articles, il existe des cas dans lesquels les tests OPG IM rejettent les hypothèses nulles correctes presque tout le temps. Le problème semble empirer lorsque le nombre de degrés de liberté augmente, et il ne s'atténue pas substantiellement quand la taille de l'échantillon augmente. Un exemple extrême, dans Davidson et MacKinnon (1992), est un modèle de régression linéaire avec 10 régresseurs, et donc 65 degrés de liberté, pour lequel la forme OPG du test

IM rejette la véritable hypothèse nulle à un niveau nominal de 5% dans 99.9% des expériences lorsque  $n = 200$ , et 92.7% lorsque  $n = 1000$ .

Heureusement, des méthodes alternatives de calcul des tests IM sont disponibles dans de nombreux cas. Elles ont invariablement de bien meilleures propriétés en échantillon fini que la version OPG mais ne sont pas applicables partout. Différentes techniques ont été suggérées par Chesher et Spady (1991), Orme (1990a, 1990b), et Davidson et MacKinnon (1992). Dans le dernier de ces articles, nous avons exploité un résultat important dû à Chesher (1984), qui a montré que l'alternative implicite du test IM est un modèle à variation paramétrique aléatoire. Cela nous a permis explicitement de construire un test contre ce type d'alternative pour la classe des modèles pour lesquels la DLR est applicable (voir la Section 14.4). Orme (1990b) suggère des variétés alternatives des régressions à longueur double ou triple pour calculer des tests IM dans d'autres types de modèles.

L'obtention d'une statistique de test IM incompatible avec l'hypothèse nulle (ce qui doit représenter la majorité des cas si la version OPG du test a été utilisée), ne signifie pas nécessairement que nous devons abandonner le modèle testé. Cela signifie que nous devons utiliser des méthodes d'inférences beaucoup plus robustes. Dans le cas des modèles de régression, nous avons vu dans la Section 16.3 que nous pouvons réaliser des inférences correctes en présence d'hétéroscédasticité de forme inconnue en utilisant un HCCME à la place de la matrice de covariance OLS conventionnelle. Dans le cas plus général des modèles estimés par maximum de vraisemblance, une option similaire nous est offerte. Rappelons le résultat

$$\mathbf{V}^\infty(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) = \mathcal{H}^{-1}(\boldsymbol{\theta}_0)\mathcal{J}(\boldsymbol{\theta}_0)\mathcal{H}^{-1}(\boldsymbol{\theta}_0), \quad (16.73)$$

qui était originellement (8.42). Nous avons obtenu ce résultat avant de démontrer l'égalité de la matrice d'information, que nous avons utilisée pour obtenir le résultat plus simple que

$$\mathbf{V}^\infty(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) = \mathcal{J}^{-1}(\boldsymbol{\theta}_0). \quad (16.74)$$

De plus, les hypothèses utilisées pour obtenir (16.73) n'étaient pas aussi fortes que celles utilisées pour obtenir l'égalité de la matrice d'information. Cela suggère que (16.73) peut être vraie plus généralement que (16.74), et cela est en effet le cas, comme White (1982) l'a montré. Ainsi, s'il y a raison de croire que l'égalité de la matrice d'information ne tient pas, il peut être astucieux d'employer l'estimateur suivant pour la matrice de covariance de  $\hat{\boldsymbol{\theta}}$ :

$$\hat{\mathbf{H}}^{-1}(\hat{\mathbf{G}}^\top \hat{\mathbf{G}})\hat{\mathbf{H}}^{-1}, \quad (16.75)$$

où  $\hat{\mathbf{H}}$  désigne la matrice Hessienne évaluée avec les estimations ML de  $\hat{\boldsymbol{\theta}}$ . En tant qu'analogue naturel de (8.42), l'expression (16.75) sera asymptotiquement valable sous des conditions plus faibles que celles qui permettent l'utilisation de  $-\hat{\mathbf{H}}^{-1}$  ou  $(\hat{\mathbf{G}}^\top \hat{\mathbf{G}})^{-1}$ .



## 16.10 CONCLUSION

Ce chapitre a couvert un ensemble de concepts très important, dont beaucoup sont relativement récents et certains n'ont qu'un faible lien avec le thème de l'hétéroscédasticité. Le thème unificateur du chapitre concerne les moments de la variable dépendante d'ordre supérieur à un. Dans le prochain chapitre, nous continuerons de souligner le rôle des moments en introduisant une méthode d'estimation importante appelée méthode des moments généralisée. Une grande part des concepts couverts dans ce chapitre, tels que les HCCME et les tests de moments conditionnels, y réapparaîtront.

## TERMES ET CONCEPTS

ARCH-en-moyenne (ARCH-M)	moments empiriques
conditions de moment (conditionnel et non conditionnel)	test de Goldfeld-Quandt pour l'hétéroscédasticité
directions scédastiques	test de White pour l'hétéroscédasticité
estimateur de la matrice de covariance robuste à l'hétéroscédasticité (HCCME)	tests de la matrice d'information (tests IM)
hétéroscédasticité conditionnelle autorégressive (ARCH)	tests de spécification
matrice de covariance OLS généralisée	tests de moments conditionnels (tests CM)
modèle ARCH( $p$ )	tests pour l'asymétrie et l'excès de kurtosis (aplatissement)
modèle ARCH généralisé (GARCH)	
modèle GARCH( $p, q$ )	

# Chapitre 17

## La Méthode des Moments Généralisée

### 17.1 INTRODUCTION ET DÉFINITIONS

Nous avons vu au cours du chapitre précédent que si un modèle est correctement spécifié, certains moments conditionnels seront nuls. L'idée fondamentale de la **méthode des moments généralisée**, ou **GMM**, est que les conditions qui portent sur les moments peuvent être exploitées non seulement pour tester la spécification d'un modèle mais aussi pour définir les paramètres du modèle, dans le sens où elles fournissent une application définissante des paramètres pour un modèle. L'exemple de base qui illustre cette idée est celui d'un modèle pour lequel le seul paramètre qui nous intéresse est l'espérance de la variable dépendante. Ceci est un cas particulier de ce que l'on appelle un **modèle de localisation**. Si chaque observation sur une variable dépendante  $y$  est un tirage issu d'une loi de distribution d'espérance  $m$ , alors le moment  $E(y - m)$  doit être nul. Cette propriété permet de *définir* le paramètre  $m$ , puisque si  $m' \neq m$ ,  $E(y - m') \neq 0$ . Autrement dit, la condition portant sur le moment n'est satisfaite que pour la véritable valeur du paramètre.

En accord avec la **méthode des moments** (ordinaire), si l'on dispose d'un échantillon de tirages indépendants issus d'une quelconque loi de distribution, il est possible d'estimer n'importe quel moment de la distribution par le moment empirique correspondant. Cette procédure se justifie très facilement en invoquant la loi des grands nombres sous sa forme la plus simple. Ainsi, pour le modèle de localisation, si l'on note les observations  $y_t$ ,  $t = 1, \dots, n$ , l'estimateur de la méthode des moments de  $m$  correspond précisément à la moyenne empirique

$$\hat{m} = \frac{1}{n} \sum_{t=1}^n y_t. \quad (17.01)$$

Lorsque l'on évoque la méthode des moments *généralisée*, cela implique en réalité de nombreuses généralisations. Certaines n'impliquent pas plus que l'abandon de conditions de régularité, par exemple l'hypothèse d'observations i.i.d. Puisque de nombreuses lois des grands nombres différentes peuvent être démontrées (souvenons-nous de la liste donnée dans la Section 4.7), il n'y a aucune raison de se limiter aux cas où les observations sont i.i.d. Mais les généralisations fondamentales proviennent de deux éléments. Le premier est

que les moments conditionnels peuvent être utilisés également comme des moments non conditionnels, et le second est que les moments peuvent dépendre de paramètres inconnus.

C'est la seconde généralisation que nous utilisons à présent pour obtenir l'**estimateur de la méthode des moments généralisée**, ou **estimateur GMM**, de  $m$  dans le modèle de localisation. Nous oublions pour l'instant que  $m$  est lui-même un moment et utilisons la **condition portant sur le moment**

$$E(y - m) = 0 \quad (17.02)$$

pour définir  $m$ . L'essence de la méthode des moments, qu'elle soit ordinaire ou généralisée, consiste à remplacer les moments théoriques de la population par les moments empiriques. Nous remplaçons par conséquent l'espérance dans (17.02) par la moyenne empirique et définissons  $\hat{m}$  de façon implicite par

$$\frac{1}{n} \sum_{t=1}^n (y_t - \hat{m}) = 0,$$

que nous résolvons immédiatement pour obtenir le même estimateur que dans (17.01).

L'estimateur le plus fréquemment utilisé en économétrie, à savoir l'estimateur OLS, peut être considéré comme un estimateur GMM. Nous mettrons à jour plusieurs caractéristiques générales de l'estimateur GMM en l'examinant sous cet angle. Lorsque l'on écrit

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (17.03)$$

l'interprétation habituelle que l'on donne est

$$E(y_t | \Omega_t) = \mathbf{X}_t \boldsymbol{\beta} \quad \text{pour } t = 1, \dots, n, \quad (17.04)$$

où  $\Omega_t$  désigne un ensemble d'information quelconque. Ceci implique l'égalité  $E(u_t | \Omega_t) = 0$ . Bien souvent, nous formulons des hypothèses supplémentaires sur  $\mathbf{u}$ , telles que l'indépendance en série, l'homoscédasticité, ou même la normalité. Pour nos préoccupations actuelles, aucune de ces hypothèses n'est nécessaire.

Si, comme d'habitude,  $k$  désigne le nombre de paramètres dans (17.03), il est clair que nous avons besoin d'au moins  $k$  conditions portant sur les moments pour définir un ensemble complet d'estimations paramétriques. Mais (17.04) ne semble pas en fournir plus d'une. La façon de résoudre ce dilemme constitue l'une des caractéristiques majeures de la GMM. Puisque (17.04) fournit une condition portant sur le moment *conditionnel*  $E(u_t | \Omega_t) = 0$ , il s'ensuit que, pour tout vecteur  $\mathbf{w}$  tel que  $w_t \in \Omega_t$ , les moments non conditionnels  $E(w_t(y_t - \mathbf{X}_t \boldsymbol{\beta}))$  sont nuls. De façon minimale, les régresseurs  $\mathbf{X}_t$  appartiennent à l'ensemble d'informations  $\Omega_t$ , et il y en a précisément  $k$ . Nous

pouvons donc utiliser les  $k$  régresseurs pour définir les  $k$  conditions portant sur les moments non conditionnels. La contrepartie empirique de ces conditions est donnée par le vecteur colonne

$$\frac{1}{n} \sum_{t=1}^n \mathbf{X}_t^\top (y_t - \mathbf{X}_t \boldsymbol{\beta}).$$

Il est clair que l'on obtient les conditions du premier ordre (1.03) utiles à la définition de l'estimateur OLS en annulant ces conditions portant sur les moments empiriques. Il apparaît, par la suite, que l'estimateur OLS, en tant qu'estimateur GMM, devrait être applicable sans aucune des hypothèses que l'on formule généralement sur les moments d'ordre deux des aléas, telles que l'indépendance en série ou l'homoscédasticité, et qui influencent la structure de leur matrice de variance-covariance. En réalité, la *convergence* de l'estimateur OLS ne provient que du fait que cet estimateur satisfait certaines conditions portant sur les moments. Cela viendra de la démonstration de la convergence de l'estimateur GMM que nous développerons dans section suivante, bien que cela paraisse naturel.

On peut dériver l'estimateur simple des variables instrumentales (7.25) de la même manière que l'estimateur OLS. L'éventuelle endogénéité des régresseurs  $\mathbf{X}$  dans (17.03) peut signifier que nous *ne voulons pas* imposer la condition  $E(u_t | \Omega_t) = 0$ . Cependant, nous réclamons, soit par une connaissance a priori soit par hypothèse, qu'il existe une matrice  $\mathbf{W}$  de dimension  $n \times k$  d'instruments valables, avec une ligne type  $\mathbf{W}_t \in \Omega_t$ . Ceci implique que nous pouvons utiliser les  $k$  conditions portant sur les moments  $E(\mathbf{W}_t u_t) = \mathbf{0}$ . Les contreparties empiriques de ces conditions sont

$$\frac{1}{n} \sum_{t=1}^n \mathbf{W}_t^\top (y_t - \mathbf{X}_t \boldsymbol{\beta}) = \mathbf{0}$$

ou, en omettant le facteur  $n^{-1}$  et en utilisant une notation matricielle,

$$\mathbf{W}^\top (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}) = \mathbf{0}. \quad (17.05)$$

Ces équations correspondent aux conditions du premier ordre qui définissent un estimateur IV simple.

Les deux exemples précédents montrent que les variables instrumentales, et parmi elles les régresseurs utilisés comme instruments, génèrent des conditions sur les moments comme celles employées dans les tests de spécification du moment conditionnel de la Section 16.8. De même que les conditions sur les moments peuvent avoir de nombreuses sources, les variables instrumentales de nombreuses sortes peuvent se suggérer d'elles-mêmes dans le contexte d'un quelconque modèle économétrique donné. Il en résulte qu'il y a habituellement beaucoup plus d'instruments disponibles que nous n'en avons besoin

pour identifier les paramètres du modèle. Souvenons-nous que, dans le contexte de la régression linéaire (17.03), *tout* vecteur  $\mathbf{w}$  tel que  $w_t \in \Omega_t$  peut être employé. Ces instruments gratuits, comme nous allons le voir dans peu de temps, peuvent être exploités dans le contexte de la GMM, tout comme ils le sont dans le contexte des IV, pour générer des contraintes de suridentification qui peuvent avoir un double rôle: améliorer l'efficacité des estimations des paramètres et tester la spécification du modèle.

L'estimation GMM n'est bien évidemment pas limitée aux modèles de régression linéaire. Nous allons à présent établir certaines définitions dans un contexte non linéaire plus général, mais qui reste encore relativement simple. Nous nous limitons par conséquent temporairement au cas des modèles juste identifiés. Le cas plus réaliste des modèles suridentifiés sera l'objet de la section suivante.

Notre première tâche consiste à caractériser d'une manière quelconque des modèles que l'on espère estimer par GMM. Dans le Chapitre 5, nous définissions un modèle économétrique comme un ensemble de DGP. Un modèle paramétrique était défini comme un modèle associé à une **application définissante des paramètres**, qui associe un vecteur de paramètres appartenant à un espace paramétrique quelconque à chaque DGP du modèle. Dans le contexte de la GMM, il existe de nombreuses façons possibles de choisir le modèle, c'est-à-dire l'ensemble des DGP. L'un des avantages de la GMM en tant que méthode d'estimation est qu'elle permet la manipulation de modèles composés d'un très grand nombre de DGP. En nette opposition avec l'estimation ML, où le modèle doit être spécifié totalement, tout DGP est admissible s'il satisfait un petit nombre de contraintes ou de conditions de régularité. Quelquefois, seule l'existence des moments utilisés pour définir les paramètres est requise pour qu'un modèle soit bien défini. Quelquefois, le chercheur souhaitera imposer une structure plus complète au modèle, éliminant des DGP qui auraient sinon été contenus dans le modèle. Cela pourra se faire en formulant des hypothèses telles que l'homoscédasticité ou l'indépendance en série, ou encore l'existence de moments autres que ceux qui définissent les paramètres. Notre préoccupation immédiate consiste à détailler simplement la spécification du modèle, aussi supposons-nous simplement qu'un ensemble de DGP  $\mathcal{M}$  a été choisi pour représenter le modèle.

L'exigence suivante concerne l'application définissante des paramètres. Ce sont les conditions portant sur les moments qui y pourvoient, puisqu'elles fournissent une définition *implicite* de l'application. Notons  $f_{ti}(y_t, \boldsymbol{\theta})$ ,  $i = 1, \dots, k$ , une fonction de la variable dépendante ou d'un vecteur de variables dépendantes  $y_t$ . Nous supposons que cette fonction possède une espérance nulle pour tout DGP du modèle caractérisé par le vecteur des paramètres  $\boldsymbol{\theta}$  de dimension  $k$ . En général, parce que toute la théorie de ce chapitre est asymptotique,  $t$ , qui est l'indice des observations, peut prendre n'importe quelle valeurentière positive. Dans la pratique, les fonctions  $f_{ti}$  dépendront fréquemment des variables exogènes et prédéterminées ainsi que de la (des)

variable(s) dépendante(s). Ainsi les conditions sur les moments

$$E(f_{ti}(y_t, \boldsymbol{\theta})) = 0, \quad i = 1, \dots, k, \quad (17.06)$$

fournissent une application définissante des paramètres sous des conditions de régularité adéquates. Ces conditions assurent que, pour chaque DGP  $\mu$  appartenant au modèle  $\mathbb{M}$ , il n'existe qu'un seul vecteur de paramètres  $\boldsymbol{\theta}$  d'un espace paramétrique quelconque  $\Theta$  qui annule les espérances (17.06). Il est généralement commode d'exiger en plus que, pour tous les DGP dans le modèle, et pour tout vecteur  $\boldsymbol{\theta} \in \Theta$ , les espérances dans (17.06) existent.

Comme c'est le cas avec tous les autres modèles paramétriques considérés jusqu'à présent, l'existence d'une application définissante des paramètres bien définie garantit l'identification asymptotique des paramètres du modèle. Leur identification par un échantillon donné dépend de l'existence d'une unique solution à ce que l'on pourrait appeler des **équations définissantes des paramètres** qui sont les contreparties empiriques des conditions portant sur les moments (17.06). Ces équations définissantes de l'estimateur, qui annulent les **moments empiriques**, sont

$$\frac{1}{n} \sum_{t=1}^n f_{ti}(y_t, \boldsymbol{\theta}) = 0, \quad i = 1, \dots, k. \quad (17.07)$$

S'il existe un unique vecteur  $\hat{\boldsymbol{\theta}}$  qui satisfait (17.07), alors le modèle est identifié par les données et  $\hat{\boldsymbol{\theta}}$  est, par définition, l'estimateur GMM de  $\boldsymbol{\theta}$ .

La méthode des moments généralisée fut suggérée sous cette appellation par Hansen (1982), mais l'idée de base remonte au moins à Sargan (1958). Un cas particulier de la GMM appelé doubles moindres carrés en deux étapes fut proposé par Cumby, Huizinga, et Obstfeld (1983). L'une des motivations au développement de la méthode était l'intérêt croissant durant le début des années 80 pour les modèles d'anticipations rationnelles. Un principe fondamental de ces modèles est que les erreurs d'anticipations doivent être indépendantes de toutes les variables des ensembles d'information des agents qui forment ces anticipations. Par conséquent, les erreurs de prévision, les échecs à atteindre un optimum, et d'autres conséquences (mesurables) de prévision imparfaite doivent être, si les anticipations sont véritablement formulées de façon rationnelle, indépendantes des variables appartenant aux ensembles d'information individuels au moment où les anticipations se forment. Cette indépendance fait apparaître des conditions variées sur les moments conditionnels, qui donnent lieu par la suite à des conditions sur les moments (non conditionnels) sur lesquels on peut fonder l'estimation GMM. La première application importante de cette idée apparaît chez Hansen et Singleton (1982), qui utilisent les conditions stochastiques d'Euler associées aux problèmes d'optimisation intertemporelle des agents en tant que source de leurs conditions sur les moments conditionnels. D'autres applications de

la GMM se trouvent chez Dunn et Singleton (1986), Eichenbaum, Hansen, et Singleton (1988), et Epstein et Zin (1991).

Nous avons esquissé à présent la plupart des résultats importants relatifs à l'estimation GMM. Il reste à considérer la manière de traiter les conditions de suridentification, d'exhiber les propriétés théoriques des estimateurs GMM, de savoir comment calculer au mieux les estimations GMM dans la pratique, et de trouver des procédures de test comparables aux tests du moment conditionnel dans un contexte GMM. Dans la section qui suit, nous discutons de la théorie asymptotique de ce que l'on appelle les **M-estimateurs**, c'est-à-dire des estimateurs définis par la maximisation ou la minimisation d'une fonction critère quelconque. Nous établissons le lien entre ces estimateurs et les estimations GMM et étudions brièvement les conditions de régularité. Puis, dans la Section 17.3, nous portons notre attention sur les questions d'efficacité et d'inférence, dans un traitement simultané puisque toutes deux dépendent de la matrice de covariance asymptotique des estimations des paramètres. Ces thèmes sont également discutés dans la Section 17.4, dont le thème principal est le choix des instruments et des conditions sur les moments. La Section 17.5 nous donnera l'occasion de discuter du problème pratique de l'estimation de la matrice de covariance. Cette discussion est plus délicate pour la GMM que pour de nombreuses autres techniques, parce que la GMM affecte la matrice de pondération que l'on utilise dans la fonction critère. Enfin, dans la Section 17.6, nous discutons des tests de spécification dans le contexte de l'estimation GMM.

## 17.2 FONCTIONS CRITÈRE ET M-ESTIMATEURS

Dans le Chapitre 7, l'estimateur IV pour le modèle de régression linéaire a été défini par la minimisation de la **fonction critère**

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{P}_W (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}); \quad (17.08)$$

voir l'équation (7.15). Notons  $k$  le nombre des régresseurs et  $l \geq k$  le nombre des instruments. Dans le cas juste identifié, pour lequel  $l = k$ , la valeur de la fonction critère minimisée est nulle. Cette valeur de la fonction est atteinte lorsque la valeur  $\boldsymbol{\beta}$  est donnée par l'estimateur IV simple, défini par les  $k$  conditions (17.05). Lorsque  $l > k$ , la valeur minimisée est en général strictement positive, puisqu'il n'est pas possible en général de résoudre ce qui est désormais un ensemble de  $l$  conditions (17.05) pour  $k$  inconnues.

Le cas suridentifié dans le contexte de la GMM est similaire. Il y a  $l$  équations définissantes de l'estimateur (17.07) mais seulement  $k$  inconnues. Au lieu de résoudre un ensemble d'équations, nous allons utiliser les membres de gauche de ces équations pour définir une fonction critère qui est par conséquent minimisée pour fournir les estimations des paramètres. Considérons à nouveau (17.08). Si nous l'écrivons sous la forme

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (17.09)$$

nous observons que l'expression est une forme quadratique composée des moments empiriques  $\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$  et de l'inverse de la matrice définie positive  $\mathbf{W}^\top\mathbf{W}$ . Cette matrice définie positive est, sous les hypothèses d'homoscédasticité et d'indépendance en série, proportionnelle à la matrice de covariance du vecteur des moments, le facteur de proportionnalité étant la variance des aléas. L'omission de ce facteur de proportionnalité importe peu, parce que la valeur de  $\boldsymbol{\beta}$  qui minimise (17.09) est inchangée si (17.09) est multipliée par n'importe quelle valeur scalaire positive.

Il n'est pas utile d'employer la matrice de covariance des moments empiriques  $\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$  si l'on veut seulement obtenir des estimations *convergentes*, plutôt qu'*efficaces*, de  $\boldsymbol{\beta}$  par la minimisation de la fonction critère. Si nous remplaçons  $(\mathbf{W}^\top\mathbf{W})^{-1}$  dans (17.09) par n'importe quelle matrice  $\mathbf{A}(\mathbf{y})$  asymptotiquement déterministe, symétrique, définie positive et de dimension  $l \times l$ , la fonction critère devient

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}\mathbf{A}(\mathbf{y})\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (17.10)$$

et nous voyons aisément que l'estimateur qui en découle est

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W}\mathbf{A}(\mathbf{y})\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{A}(\mathbf{y})\mathbf{W}^\top \mathbf{y}.$$

Si  $l = k$  et si la matrice  $\mathbf{W}^\top \mathbf{X}$  est carrée et non singulière, cette expression se réduit à l'estimateur IV simple  $(\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{y}$ , quel que soit le choix de  $\mathbf{A}$ . Le choix de  $\mathbf{A}$  est sans conséquence dans ce cas parce que le nombre des conditions sur les moments est égal au nombre des paramètres, ce qui implique que (17.10) atteint toujours un minimum égal à zéro pour toute matrice  $\mathbf{A}$ .

En général, si  $\mathbf{W}$  est une matrice d'instruments valables,  $\hat{\boldsymbol{\beta}}$  sera un estimateur convergent de  $\boldsymbol{\beta}$ , comme nous le constatons à l'aide d'arguments standards. Sous les hypothèses d'homoscédasticité et d'indépendance en série des aléas, l'estimateur  $\hat{\boldsymbol{\beta}}$  est malgré tout moins efficace que l'estimateur IV habituel  $\tilde{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y}$ , à moins que  $\mathbf{A}$  ne soit proportionnelle à  $(\mathbf{W}^\top \mathbf{W})^{-1}$ . La démonstration de ce résultat est similaire aux démonstrations du Théorème de Gauss-Markov (Théorème 5.3) et de la borne inférieure de Cramér-Rao dans la Section 8.8. Nous démontrons que la différence  $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$  est asymptotiquement non corrélée à  $\tilde{\boldsymbol{\beta}}$ . Cela implique que la matrice de covariance asymptotique de  $\hat{\boldsymbol{\beta}}$  est la somme des matrices de covariance asymptotique de  $\tilde{\boldsymbol{\beta}}$  et de la différence entre les deux estimateurs. Par conséquent,  $\hat{\boldsymbol{\beta}}$  doit être au moins aussi efficace que  $\tilde{\boldsymbol{\beta}}$ . La différence entre les deux estimateurs est

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{W}\mathbf{A}\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{A}\mathbf{W}^\top \mathbf{y} - (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{W}\mathbf{A}\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{A}\mathbf{W}^\top \mathbf{M}_X^W \mathbf{y}, \end{aligned} \quad (17.11)$$

où la matrice de projection oblique  $\mathbf{M}_X^W$  est définie par

$$\mathbf{M}_X^W = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W.$$



La construction de (17.11) n'a pas été détaillée totalement, parce qu'elle est essentiellement la même que les nombreuses précédentes; voir, par exemple, (7.59).

Puisque  $\mathbf{M}_X^W \mathbf{X} = \mathbf{0}$ , nous pouvons remplacer  $\mathbf{y}$  dans l'expression (17.11) par  $\mathbf{u}$  si  $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{u}$  pour un quelconque vecteur  $\beta_0$ . Il est désormais possible de voir que  $\tilde{\beta}$  est asymptotiquement non corrélé à (17.11). La partie aléatoire de  $\tilde{\beta}$  est  $\mathbf{X}^\top \mathbf{P}_W \mathbf{u}$ , et la partie aléatoire de (17.11) est  $\mathbf{W}^\top \mathbf{M}_X^W \mathbf{u}$ . Lorsque les aléas sont homoscedastiques, indépendants en série et ont une variance égale à  $\sigma^2$ , la matrice des covariances asymptotiques de ces parties aléatoires est

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sigma^2 \mathbf{X}^\top \mathbf{P}_W (\mathbf{M}_X^W)^\top \mathbf{W} \right).$$

Or cette matrice est nulle, comme nous le démontrons, puisque

$$\mathbf{X}^\top \mathbf{P}_W (\mathbf{M}_X^W)^\top \mathbf{W} = \mathbf{X}^\top \mathbf{W} - \mathbf{X}^\top \mathbf{P}_W \mathbf{X} (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} = \mathbf{0}.$$

Dans la prochaine section, nous discuterons ce résultat plus finement. Il confère simplement une sorte d'optimalité ou d'efficacité à l'estimateur IV habituel, et il sera intéressant d'étudier la nature exacte de cette optimalité.

Dans le contexte plus général de la GMM, nous pouvons construire une fonction critère à des fins d'estimation en utilisant une matrice  $\mathbf{A}(\mathbf{y})$  arbitrairement symétrique, définie positive, éventuellement dépendante des données, et  $O(1)$ . Nous appellerons  $\mathbf{A}$  **matrice de pondération** et exigerons que, pour chaque DGP  $\mu$  appartenant au modèle  $\mathbb{M}$ ,

$$\text{plim}_{n \rightarrow \infty} \mu \mathbf{A}(\mathbf{y}) = \mathbf{A}_0(\mu), \quad (17.12)$$

où  $\mathbf{A}_0(\mu)$  est une matrice finie, déterministe, symétrique et définie positive. Notons  $\mathbf{F}(\mathbf{y}, \boldsymbol{\theta})$  la matrice dont l'élément type est  $f_{ti}(y_t, \boldsymbol{\theta})$  où, comme pour (17.07),  $f_{ti}(y_t, \boldsymbol{\theta})$  désigne la contribution de l'observation  $t$  au  $i^{\text{ième}}$  moment. Nous supposons que  $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$  et que  $1 \leq i \leq l$ , avec  $l > k$ . Alors, si  $\boldsymbol{\iota}$ , comme d'habitude, désigne le vecteur de dimension  $n$  dont chaque composante est égale à 1, les conditions sur les moments empiriques sont données par

$$\mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta}) \boldsymbol{\iota} = \mathbf{0},$$

et une fonction critère admissible pour estimer  $\boldsymbol{\theta}$  est

$$\boldsymbol{\iota}^\top \mathbf{F}(\mathbf{y}, \boldsymbol{\theta}) \mathbf{A}(\mathbf{y}) \mathbf{F}^\top(\mathbf{y}, \boldsymbol{\theta}) \boldsymbol{\iota}. \quad (17.13)$$

Nous établissons à présent le résultat fondamental pour montrer que l'estimateur  $\hat{\boldsymbol{\theta}}$  issu de la minimisation de (17.13) est convergent sous certaines conditions de régularité. Ce résultat indique que si un échantillon est

généralisé par le DGP  $\mu \in \mathbb{M}$ , le véritable vecteur de paramètres  $\theta(\mu)$  minimise la limite en probabilité de  $n^{-2}$  fois la fonction critère (17.13):

$$\theta(\mu) = \operatorname{argmin}_{\theta \in \Theta} \left( \operatorname{plim}_{\mu} \left( n^{-2} \boldsymbol{\iota}^\top \mathbf{F}(\mathbf{y}, \theta) \mathbf{A}(\mathbf{y}) \mathbf{F}^\top(\mathbf{y}, \theta) \boldsymbol{\iota} \right) \right). \quad (17.14)$$

La notation  $\operatorname{plim}_{\mu}$  implique que le DGP utilisé pour calculer la limite en probabilité est  $\mu$ , et (17.14) implique que cette limite en probabilité est déterministe. Le facteur inhabituel  $n^{-2}$  apparaît parce que nous avons supposé que la matrice de pondération limite  $\mathbf{A}_0(\mu)$  est  $O(1)$ . Puisque nous nous attendons à ce que  $\mathbf{F}^\top \boldsymbol{\iota}$  soit  $O(n)$ , nous avons besoin de deux facteurs de  $n^{-1}$  pour que (17.14) soit  $O(1)$  lorsque  $n \rightarrow \infty$ .

Pour que le résultat (17.14) soit vrai, nous devons être capables d'appliquer une loi des grands nombres à  $n^{-1} \mathbf{F}^\top \boldsymbol{\iota} = n^{-1} \sum_{t=1}^n \mathbf{F}_t^\top$ , où  $\mathbf{F}_t$  est la  $t^{\text{ième}}$  ligne de  $\mathbf{F}$ . Puisque  $\mathbf{F}$  dépend de paramètres, la loi des grands nombres doit s'appliquer de façon uniforme par rapport à ces paramètres, aussi supposons-nous simplement que la condition WULLN donnée dans la Définition 4.17 s'applique à chaque composante de la série  $\{\mathbf{F}_t^\top(\theta)\}$  au moins en un voisinage quelconque du véritable vecteur de paramètres  $\theta_0 \equiv \theta(\mu)$ . Cela nous permet de poser la définition suivante:

$$\mathbf{m}(\mu, \theta) = \operatorname{plim}_{\mu} \left( \frac{1}{n} \mathbf{F}^\top(\theta) \boldsymbol{\iota} \right) = \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n E_{\mu}(\mathbf{F}_t(\theta)) \right). \quad (17.15)$$

Les conditions sur les moments empiriques (17.06) en jonction avec l'exigence que ces conditions identifient les paramètres garantissent que

$$\mathbf{m}(\mu, \theta_0) = \mathbf{0} \quad \text{et} \quad \mathbf{m}(\mu, \theta) \neq \mathbf{0} \quad \text{si} \quad \theta \neq \theta_0. \quad (17.16)$$

Puisque  $\operatorname{plim}_{\mu} \mathbf{A}(\mathbf{y}) = \mathbf{A}_0(\mu)$ , il s'ensuit que

$$\operatorname{plim}_{\mu} \left( n^{-2} \boldsymbol{\iota}^\top \mathbf{F}(\mathbf{y}, \theta) \mathbf{A}(\mathbf{y}) \mathbf{F}^\top(\mathbf{y}, \theta) \boldsymbol{\iota} \right) = \mathbf{m}^\top(\mu, \theta) \mathbf{A}_0(\mu) \mathbf{m}(\mu, \theta).$$

Puisque  $\mathbf{A}_0(\mu)$  est définie positive, cette expression est nulle pour  $\theta = \theta_0$  et (strictement) positive sinon. Cela établit (17.14).

Le résultat (17.14) implique que l'estimateur de  $\theta$  obtenu en minimisant la fonction critère (17.13) est convergent, en vertu des mêmes arguments utilisés dans les Chapitres 5 et 8 pour montrer la convergence des estimateurs NLS et ML. Comme dans le Chapitre 8, pour qu'un modèle GMM soit asymptotiquement identifié sur un espace paramétrique non compact, nous devons supposer qu'il n'existe aucune série de vecteurs de paramètres sans point limite telle que (17.13) évaluée en des points de la série tende supérieurement vers la valeur de (17.13) au véritable vecteur de paramètres  $\theta_0$ ; souvenons-nous de la Définition 8.1.

Il est pratique à cette étape d'abandonner un cas spécifique de la GMM et de traiter le problème plus général des  $M$ -estimateurs. Cette terminologie naquit dans la littérature de l'estimation robuste—voir Huber (1972, 1981)—mais en économétrie elle est souvent utilisée pour faire référence à n'importe quel estimateur associé à la maximisation ou la minimisation d'une fonction critère. Ces dernières années, un effort substantiel s'est porté sur le développement d'une théorie unifiée de tous les estimateurs de ce type. L'article qui marque une étape décisive est celui de Burguete, Gallant, et Souza (1982). Notre traitement sera relativement élémentaire; pour compléter les notions, les lecteurs devraient consulter Bates et White (1985), Gallant (1987), ou Gallant et White (1988).

Il nous faut tout d'abord poser certaines définitions. Supposons que nous travaillons avec un modèle paramétrique  $(\mathbb{M}, \boldsymbol{\theta})$ . L'espace d'arrivée de l'application définissante des paramètres  $\boldsymbol{\theta}$  sera l'espace paramétrique  $\Theta \in \mathbb{R}^k$ . Soit  $Q^n(\mathbf{y}^n, \boldsymbol{\theta})$  la valeur d'une fonction critère, où  $\mathbf{y}^n$  est un échantillon comportant  $n$  observations sur une ou plusieurs variables dépendantes, et où  $\boldsymbol{\theta} \in \Theta$ . Notons que, par un léger abus de notation,  $\boldsymbol{\theta}$  désigne à la fois l'application définissante des paramètres et les valeurs de l'application. A proprement parler, nous devrions faire référence à  $\boldsymbol{\theta}(\mu)$  pour le vecteur de paramètres associé au DGP  $\mu \in \mathbb{M}$ , mais il est inutile en général de spécifier  $\mu$  explicitement. Habituellement,  $Q^n$  dépendra autant des variables exogènes et prédéterminées que de la (des) variable(s) dépendante(s)  $\mathbf{y}^n$ . Alors, pour que la série  $Q \equiv \{Q^n\}$  soit appropriée à l'estimation des paramètres  $\boldsymbol{\theta}$ , nous exigeons que  $Q$  *identifie* ces paramètres, dans le sens de la Définition 17.1:

*Définition 17.1.*

Une série de fonctions critère  $Q$  identifie asymptotiquement un modèle paramétrique  $(\mathbb{M}, \boldsymbol{\theta})$  si, pour tout  $\mu \in \mathbb{M}$  et pour tout  $\boldsymbol{\theta} \in \Theta$ ,

$$\bar{Q}(\mu, \boldsymbol{\theta}) \equiv \text{plim}_{n \rightarrow \infty} Q^n(\mathbf{y}^n, \boldsymbol{\theta})$$

existe et satisfait l'inégalité  $\bar{Q}(\mu, \boldsymbol{\theta}(\mu)) < \bar{Q}(\mu, \boldsymbol{\theta})$  pour tout vecteur de paramètres  $\boldsymbol{\theta} \neq \boldsymbol{\theta}(\mu)$ . En plus de cela, si  $\Theta$  est non compact, il n'existe aucune série  $\{\boldsymbol{\theta}^m\}$  sans point limite telle que

$$\lim_{m \rightarrow \infty} \bar{Q}(\mu, \boldsymbol{\theta}^m) = \bar{Q}(\mu, \boldsymbol{\theta}(\mu)).$$

Alors, bien que nous présentions une démonstration peu rigoureuse, nous voyons intuitivement que l'estimateur  $\hat{\boldsymbol{\theta}}_Q \equiv \{\hat{\boldsymbol{\theta}}_Q^n\}$  défini par

$$\hat{\boldsymbol{\theta}}_Q^n = \underset{\boldsymbol{\theta} \in \Theta}{\text{argmin}} Q^n(\mathbf{y}^n, \boldsymbol{\theta}) \quad (17.17)$$

devrait converger vers  $\boldsymbol{\theta}$ , c'est-à-dire,

$$\text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\theta}}_Q^n = \boldsymbol{\theta}(\mu). \quad (17.18)$$

Une démonstration peu rigoureuse de (17.18) emploie exactement les mêmes arguments que ceux employés dans la Section 8.4, et qui menaient à l'équation (8.31). Le résultat formel peut s'énoncer ainsi:

*Théorème 17.1. Convergence des M-Estimateurs*

Le  $M$ -estimateur défini par la minimisation de la série des fonctions critère  $Q$  converge vers les paramètres d'un modèle paramétrique  $(\mathbb{M}, \boldsymbol{\theta})$  si la série  $Q$  identifie le modèle au sens de la Définition 17.1.

La fait que  $Q^n(\boldsymbol{\theta}) = O(1)$  lorsque  $n \rightarrow \infty$  est implicite dans la Définition 17.1. Ainsi la plupart des fonctions critère qui sont en réalité utilisées devront être multipliées par des puissances de  $n$  avant de savoir si elles vérifient la Définition 17.1. La fonction somme-des-carrés utilisée dans l'estimation NLS et la fonction de logvraisemblance utilisée dans l'estimation ML, par exemple, sont toutes deux  $O(n)$  et doivent donc être divisées par  $n$ , comme dans les équations (5.10) et (8.31). Puisque nous avons supposé dans (17.12) que  $\mathbf{A}$  est  $O(1)$ , la fonction critère (17.13) doit être divisée par  $n^2$ , comme nous l'avons déjà mentionné dans (17.14).

La convergence du  $M$ -estimateur (17.17) étant établie, il est temps de passer à la normalité asymptotique. Comme toujours, cette propriété nécessite que des conditions de régularité supplémentaires soient satisfaites. Jusqu'ici, nous n'avons posé aucune hypothèse particulière sur la forme de la fonction critère  $Q^n$ . La fonction somme-des-carrés et la fonction de logvraisemblance peuvent toutes deux s'exprimer comme la somme de  $n$  contributions, une pour chaque observation de l'échantillon. La fonction critère de la GMM (17.13) adopte une structure légèrement plus compliquée: c'est une forme quadratique composée d'une matrice définie positive et d'un vecteur  $\mathbf{F}^\top \boldsymbol{\iota}$  dont chaque composante est une somme de contributions.

La première exigence supplémentaire est que le  $M$ -estimateur que l'on étudie soit, selon la terminologie du Chapitre 8, de **Type 2**, c'est-à-dire qu'il soit une solution aux conditions de premier ordre pour un minimum intérieur de la fonction critère  $Q$ . En faisant abstraction de la dépendance explicite de  $\hat{\boldsymbol{\theta}}$  à  $n$  et  $Q$  et de celle de  $Q$  à  $n$ , nous pouvons écrire les conditions de premier ordre sous la forme

$$\frac{\partial Q}{\partial \theta_j}(\hat{\boldsymbol{\theta}}) = 0 \quad \text{pour } j = 1, \dots, k. \quad (17.19)$$

Puisque  $\hat{\boldsymbol{\theta}}$  est convergent si  $Q$  identifie  $\boldsymbol{\theta}$ , il est naturel de calculer un développement en série de Taylor des conditions (17.19) autour de  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ . Cela donne

$$\frac{\partial Q}{\partial \theta_j}(\boldsymbol{\theta}_0) + \sum_{i=1}^k \frac{\partial^2 Q}{\partial \theta_j \partial \theta_i}(\boldsymbol{\theta}_j^*)(\hat{\theta}_i - \theta_i^0) = 0, \quad \text{pour } j = 1, \dots, k, \quad (17.20)$$

où  $\boldsymbol{\theta}_j^*$  est une combinaison convexe de  $\boldsymbol{\theta}_0$  et de  $\hat{\boldsymbol{\theta}}$ . Alors, à condition que la matrice Hessienne  $\mathcal{H}(\boldsymbol{\theta})$ , dont l'élément type est  $\partial^2 Q(\boldsymbol{\theta})/\partial \theta_j \partial \theta_i$ , soit inversible

au voisinage de  $\boldsymbol{\theta}_0$ , nous obtenons

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = -(\mathcal{H}^*)^{-1} \mathbf{g}(\boldsymbol{\theta}_0), \quad (17.21)$$

où  $\mathbf{g}(\boldsymbol{\theta})$  désigne le gradient de  $Q$ , c'est-à-dire le vecteur de dimension  $k$  dont la composante type est  $\partial Q(\boldsymbol{\theta})/\partial \theta_j$ . Comme d'habitude,  $\mathcal{H}^*$  désigne la matrice dont les éléments sont évalués avec le vecteur approprié  $\boldsymbol{\theta}_j^*$ .

Si nous voulons être capables de déduire la normalité asymptotique de  $\hat{\boldsymbol{\theta}}$  à partir de (17.21), il doit être possible d'appliquer une loi des grands nombres à  $\mathcal{H}^*$  et un théorème de la limite centrale à  $n^{1/2} \mathbf{g}(\boldsymbol{\theta}_0)$ . Nous obtiendrons alors le résultat suivant:

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\left(\text{plim}_{n \rightarrow \infty} \mathcal{H}_0\right)^{-1} n^{1/2} \mathbf{g}(\boldsymbol{\theta}_0). \quad (17.22)$$

De quelles conditions de régularité avons-nous besoin pour (17.22)? Il faut tout d'abord, afin de justifier le développement en série de Taylor dans (17.20), que  $Q$  soit au moins deux fois continûment différentiable par rapport à  $\boldsymbol{\theta}$ . Si c'est le cas, alors la matrice Hessienne de  $Q$  est  $O(1)$  lorsque  $n \rightarrow \infty$ . A cause de cela, nous la notons  $\mathcal{H}_0$  plutôt que  $\mathbf{H}$ ; voir la Section 8.2. Ensuite nous avons besoin de conditions qui permettent l'application d'une loi des grands nombres et d'un théorème de la limite centrale. De façon assez formelle, nous pouvons énoncer un théorème basé sur le Théorème 8.3 comme suit:

*Théorème 17.2. Normalité Asymptotique des M-Estimateurs*

Le  $M$ -estimateur issu de la série des fonctions critère  $Q$  est asymptotiquement normal s'il satisfait les conditions du Théorème 17.1 et si de plus

- (i) pour tout  $n$  et tout  $\boldsymbol{\theta} \in \Theta$ ,  $Q^n(\mathbf{y}^n, \boldsymbol{\theta})$  est deux fois continûment différentiable par rapport à  $\boldsymbol{\theta}$  pour presque tout  $\mathbf{y}$ , et la fonction limite  $\bar{Q}(\mu, \boldsymbol{\theta})$  est deux fois continûment différentiable par rapport à  $\boldsymbol{\theta}$  pour tout  $\boldsymbol{\theta} \in \Theta$  et pour tout  $\mu \in \mathbb{M}$ ;
- (ii) pour tout DGP  $\mu \in \mathbb{M}$  et pour toute série  $\{\boldsymbol{\theta}^n\}$  qui tend en probabilité vers  $\boldsymbol{\theta}(\mu)$  lorsque  $n \rightarrow \infty$ , la matrice Hessienne  $\mathcal{H}^n(\mathbf{y}^n, \boldsymbol{\theta}^n)$  de  $Q^n$  par rapport à  $\boldsymbol{\theta}$  tend uniformément en probabilité vers une matrice  $\mathcal{H}(\mu)$  définie positive, finie et déterministe; et
- (iii) pour tout DGP  $\mu \in \mathbb{M}$ ,  $n^{1/2}$  fois le gradient de  $Q^n(\mathbf{y}^n, \boldsymbol{\theta})$ , ou  $n^{1/2} \mathbf{g}(\mathbf{y}^n, \boldsymbol{\theta}(\mu))$ , converge en distribution lorsque  $n \rightarrow \infty$  vers une distribution normale multivariée d'espérance nulle et de matrice de covariance  $\mathbf{V}(\mu)$ .

Sous ces conditions, la distribution de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}(\mu))$  tend vers  $N(0, \mathcal{H}(\mu)^{-1} \mathbf{V}(\mu) \mathcal{H}(\mu)^{-1})$ .

Il est inutile de s'attarder sur la démonstration du Théorème 17.2. Au lieu de cela, nous devrions nous ramener au cas de la GMM et chercher les conditions sous lesquelles la fonction critère (17.13), préalablement divisée par  $n^2$ ,

satisfait les exigences du théorème. Sans plus de cérémonie, nous supposons que toutes les contributions  $f_{ti}(y_t, \boldsymbol{\theta})$  sont au moins deux fois continûment différentiables par rapport à  $\boldsymbol{\theta}$  pour tout  $\boldsymbol{\theta} \in \Theta$ , pour tout  $y_t$ , et pour toutes les valeurs admissibles de n'importe quelle variable prédéterminée et exogène dont elles peuvent dépendre. Puis, nous supposons que les séries

$$\frac{1}{n} \sum_{t=1}^n \frac{\partial f_{ti}}{\partial \theta_j}(y_t, \boldsymbol{\theta}) \quad \text{et} \quad \frac{1}{n} \sum_{t=1}^n \frac{\partial^2 f_{ti}}{\partial \theta_j \partial \theta_m}(y_t, \boldsymbol{\theta})$$

pour  $i = 1, \dots, l$  et  $j, m = 1, \dots, k$  satisfont toutes deux les conditions WULLN. Cela nous permet de définir les fonctions limites comme suit:

$$d_{ij}(\mu, \boldsymbol{\theta}) \equiv \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \frac{\partial f_{ti}}{\partial \theta_j}(y_t, \boldsymbol{\theta}) \right). \quad (17.23)$$

Nous noterons  $\mathbf{D}$  la matrice de dimension  $l \times k$  dont l'élément type est  $d_{ij}$ . En rappelant la définition de  $\mathbf{m}$  dans (17.15), nous pouvons à présent affirmer que la fonction critère limite  $\bar{Q}$  empirique

$$Q^n(\mathbf{y}^n, \boldsymbol{\theta}) \equiv n^{-2} \boldsymbol{\iota}^\top \mathbf{F}(\mathbf{y}^n, \boldsymbol{\theta}) \mathbf{A}(\mathbf{y}^n) \mathbf{F}^\top(\mathbf{y}^n, \boldsymbol{\theta}) \boldsymbol{\iota} \quad (17.24)$$

est donnée par

$$\bar{Q}(\mu, \boldsymbol{\theta}) = \mathbf{m}^\top(\mu, \boldsymbol{\theta}) \mathbf{A}_0(\mu) \mathbf{m}(\mu, \boldsymbol{\theta}). \quad (17.25)$$

Bien que nous ayons supposé que les contributions  $f_{ti}$  étaient deux fois continûment différentiables, il est en général nécessaire de supposer séparément que  $\bar{Q}$  est deux fois continûment différentiable. Nous formulons donc cette hypothèse supplémentaire, qui nous permet de conclure que  $d_{ij}(\mu, \boldsymbol{\theta})$  est la dérivée de  $m_i(\mu, \boldsymbol{\theta})$ , la  $i^{\text{ième}}$  composante de  $\mathbf{m}(\mu, \boldsymbol{\theta})$ , par rapport à  $\theta_j$ . La matrice  $\mathbf{A}(\mathbf{y})$  et la matrice limite  $\mathbf{A}_0(\mu)$  ne dépendent pas du vecteur paramétrique  $\boldsymbol{\theta}$ , et nous trouvons par conséquent que le gradient de  $\bar{Q}$  par rapport à  $\boldsymbol{\theta}$  est donné par le vecteur

$$2\mathbf{D}^\top \mathbf{A}_0 \mathbf{m}. \quad (17.26)$$

A première vue, il semble qu'il n'y ait pas d'expression matricielle pratique pour la matrice Hessienne de  $\bar{Q}$ , puisque  $\mathbf{D}$  est elle-même une matrice. Cependant, lorsque  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ , nous savons à partir de (17.16) que  $\mathbf{m}(\mu, \boldsymbol{\theta}_0) = \mathbf{0}$ . Il en résulte que la matrice Hessienne limite évaluée avec le véritable vecteur de paramètres est

$$\mathcal{H}(\mu) = 2\mathbf{D}^\top(\mu, \boldsymbol{\theta}_0) \mathbf{A}_0(\mu) \mathbf{D}(\mu, \boldsymbol{\theta}_0). \quad (17.27)$$

Nous pouvons exploiter davantage les hypothèses pour garantir que les fonctions critère (17.24) et la fonction limite (17.25) satisfont les conditions (i) et (ii) du Théorème 17.2. En particulier, nous pouvons assurer que  $\mathcal{H}(\mu)$  est

définie positive du fait que  $\mathbf{D}(\mu, \boldsymbol{\theta}_0)$  devrait être de plein rang, c'est-à-dire de rang  $k$ . Cette exigence est l'analogue de l'exigence d'une **identification asymptotique forte** discutée dans le Chapitre 5 (voir le Théorème 5.2 et la discussion qui le suit), et nous adopterons une terminologie comparable dans le nouveau contexte. Cela signifie simplement que, comme les  $k$  composantes de  $\boldsymbol{\theta}$  varient au voisinage de  $\boldsymbol{\theta}_0$ , les  $l$  composantes de  $\mathbf{m}(\mu, \boldsymbol{\theta})$  varient également dans  $k$  directions indépendantes de  $\mathbb{R}^l$ .

La condition (iii) est légèrement plus délicate, puisqu'elle implique un théorème de la limite centrale. Remarquons premièrement que le gradient de  $\bar{Q}$ , évalué avec  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ , est nul, ce qui découle de (17.26). Ceci n'est qu'un reflet de la convergence de l'estimateur. Il nous faut donc remonter dans le raisonnement et considérer  $n^{1/2}$  fois le gradient de  $Q^n$  avec plus de précision. A partir de (17.24), nous obtenons, en abandonnant la dépendance explicite à la taille de l'échantillon

$$n^{1/2} \mathbf{g}_j \equiv n^{1/2} \frac{\partial Q}{\partial \theta_j} = 2 \left( \frac{1}{n} \sum_{t=1}^n \frac{\partial \mathbf{F}_t}{\partial \theta_j} \right) \mathbf{A} \left( n^{-1/2} \sum_{s=1}^n \mathbf{F}_s^\top \right), \quad (17.28)$$

où toutes les quantités sont évaluées en  $(\mathbf{y}, \boldsymbol{\theta}_0)$  et où, comme précédemment,  $\mathbf{F}_t$  est la  $t^{\text{ième}}$  ligne de  $\mathbf{F}$ . A l'évidence, notre attention doit se porter exclusivement sur le dernier facteur de l'expression,  $n^{-1/2} \sum_{s=1}^n \mathbf{F}_s^\top$ , si nous voulons obtenir la distribution asymptotique, puisque tous les autres facteurs ont de bonnes propriétés, sont déterministes, et tendent vers une limite en probabilité. Notre but n'est pas dans ce chapitre de collectionner les DGP, aussi sera-t-il suffisant pour l'instant de supposer que, pour chaque  $\mu \in \mathbb{M}$ , la série vectorielle  $\{\mathbf{F}_t(y_t, \boldsymbol{\theta}_0)\}$  obéit à la condition CLT de la Définition 4.16. C'en est assez pour la condition (iii) du Théorème 17.2, aussi pouvons-nous conclure que  $\hat{\boldsymbol{\theta}}$ , l'estimateur GMM obtenu en maximisant (17.13), est asymptotiquement normal. Remarquons que la condition CLT peut se révéler plus contraignante que ce que nous voudrions, puisqu'elle élimine certaines formes de corrélation en série; se reporter à la Section 17.5.

Il reste à calculer la matrice de covariance asymptotique de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ . Nous commençons par considérer la matrice de covariance asymptotique de (17.28),  $\mathbf{V}(\mu)$ . Soit  $\boldsymbol{\Phi}(\mu)$  une matrice de dimension  $l \times l$  définie de manière à ce que son élément type soit

$$\Phi_{ij}(\mu) \equiv \text{plim}_{\mu} \left( \frac{1}{n} \sum_{t=1}^n f_{ti}(y_t, \boldsymbol{\theta}_0) f_{tj}(y_t, \boldsymbol{\theta}_0) \right). \quad (17.29)$$

Grâce au CLT, elle correspond à la matrice de covariance asymptotique de  $n^{-1/2} \sum_{t=1}^n \mathbf{F}_t(y_t, \boldsymbol{\theta}_0)$ . Puis, étant donnée la définition (17.23), la matrice de covariance asymptotique de (17.28) est

$$\mathbf{V}(\mu) = 4 \mathbf{D}^\top(\mu, \boldsymbol{\theta}_0) \mathbf{A}_0(\mu) \boldsymbol{\Phi}(\mu) \mathbf{A}_0(\mu) \mathbf{D}(\mu, \boldsymbol{\theta}_0). \quad (17.30)$$

Par la suite, souvenons-nous qu'à partir du Théorème 17.2, la matrice de covariance asymptotique de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  est  $\mathcal{H}_0^{-1}\mathbf{V}_0\mathcal{H}_0^{-1}$ , et que, à partir de (17.27),  $\mathcal{H}_0 = 2\mathbf{D}^\top\mathbf{A}_0\mathbf{D}$ . Nous obtenons donc le résultat suivant:

$$\mathbf{V}(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)) = (\mathbf{D}^\top\mathbf{A}_0\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{A}_0\boldsymbol{\Phi}\mathbf{A}_0\mathbf{D}(\mathbf{D}^\top\mathbf{A}_0\mathbf{D})^{-1}. \quad (17.31)$$

Cette expression n'est pas particulièrement commode, bien qu'elle puisse se simplifier quelquefois, comme nous le verrons dans la section qui suit. L'estimation convergente de  $\mathbf{V}(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0))$  n'est pas difficile; il suffit d'estimer  $d_{ij}$  par

$$\frac{1}{n} \sum_{t=1}^n \frac{\partial f_{ti}}{\partial \theta_j}(\mathbf{y}, \hat{\boldsymbol{\theta}}), \quad (17.32)$$

$\mathbf{A}_0$  par  $\mathbf{A}(\mathbf{y})$ , et  $\boldsymbol{\Phi}_{ij}$  par l'expression (17.29) sans la limite en probabilité. Bien que cela fournisse une estimation convergente de (17.30), c'est souvent une estimation très parasitée. Nous parlerons de ce résultat plus en détail dans la Section 17.5, mais il est loin d'être totalement résolu.

Il est intéressant d'illustrer (17.31) dans le cas de l'estimateur IV défini par (17.08). Le résultat permettra de construire une estimation robuste à l'hétéroscédasticité de la matrice de covariance de ce dernier. Nous avons simplement besoin d'établir quelques équivalences d'ordre notationnel entre le cas IV et le cas plus général envisagé précédemment. Dans le cas IV, les éléments de la matrice  $\mathbf{F}$  deviennent  $f_{ti} = W_{ti}(y_t - \mathbf{X}_t\boldsymbol{\beta})$ . Par conséquent,

$$\mathbf{D} = -\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{W}^\top \mathbf{X} \right) \quad (17.33)$$

et

$$\mathbf{A}_0 = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{W}^\top \mathbf{W} \right)^{-1}. \quad (17.34)$$

La matrice  $\boldsymbol{\Phi}$  est obtenue à partir de (17.29):

$$\boldsymbol{\Phi} = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n (y_t - \mathbf{X}_t\boldsymbol{\beta})^2 \mathbf{W}_t^\top \mathbf{W}_t \right) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W} \right), \quad (17.35)$$

où  $\boldsymbol{\Omega}$  est la matrice diagonale dont l'élément type est  $E(y_t - \mathbf{X}_t\boldsymbol{\beta})^2$ . Si nous substituons (17.33), (17.34), et (17.35) dans (17.31), nous obtenons l'expression suivante pour la matrice de covariance asymptotique de l'estimateur IV:

$$\text{plim}_{n \rightarrow \infty} \left( \left( \frac{1}{n} \mathbf{X}^\top \mathbf{P}_W \mathbf{X} \right)^{-1} \frac{1}{n} \mathbf{X}^\top \mathbf{P}_W \boldsymbol{\Omega} \mathbf{P}_W \mathbf{X} \left( \frac{1}{n} \mathbf{X}^\top \mathbf{P}_W \mathbf{X} \right)^{-1} \right). \quad (17.36)$$

La matrice (17.36) est clairement l'analogue pour l'estimateur IV de (16.08) pour l'estimation NLS: elle fournit la matrice de covariance asymptotique



robuste à une hétéroscédasticité dont la forme est inconnue. Ainsi nous voyons que les matrices HCCME du type de celles étudiées dans la Section 16.3 sont disponibles pour l'estimateur IV. Nous pouvons alors employer n'importe quel estimateur non convergent  $\hat{\Omega}$  aperçu à cette occasion pour obtenir un estimateur convergent de  $\text{plim}(n^{-1}\mathbf{X}^\top \mathbf{P}_W \Omega \mathbf{P}_W \mathbf{X})$ .

Les lecteurs peuvent se demander à juste titre pourquoi la matrice obtenue est robuste à l'hétéroscédasticité *seulement* et non pas aussi à la corrélation en série des aléas. La réponse est que la matrice de covariance  $\mathbf{V}$  de (17.30) n'est valable que si la condition CLT est satisfaite par les contributions des moments empiriques. Celle-ci *ne sera pas* satisfaite si les aléas adoptent un schéma particulier de corrélation entre eux. Dans la Section 17.5, nous discuterons des méthodes qui permettent de traiter la corrélation en série, mais elles nous entraîneront au-delà des limites de la structure asymptotique avec laquelle nous avons travaillé jusqu'à présent.

### 17.3 ESTIMATEURS GMM EFFICACES

La question de savoir si les estimateurs GMM sont asymptotiquement efficaces n'est pas complètement directe compte tenu du fait qu'il existe de nombreux résultats distincts. Le premier résultat était dévoilé au début de la section précédente, en connexion avec l'estimation par variables instrumentales. Nous y avons vu que, pour un ensemble donné de moments empiriques  $\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ , il était possible de générer toute une famille d'estimateurs selon les choix différents de la matrice de pondération  $\mathbf{A}(\mathbf{y})$  utilisée pour construire la forme quadratique à partir des moments. Asymptotiquement, le plus efficace de ces estimateurs est obtenu en choisissant  $\mathbf{A}(\mathbf{y})$  telle qu'elle tende vers une limite en probabilité déterministe proportionnelle à l'inverse de la matrice de covariance limite des moments empiriques, préalablement pondérée par une puissance appropriée de la taille de l'échantillon  $n$ . Ce résultat revêt un caractère assez général, ainsi que nous allons le montrer.

*Théorème 17.3. Une Condition Nécessaire à l'Efficacité*

Une condition nécessaire à l'efficacité de l'estimateur issu de la minimisation de la forme quadratique (17.13) est que, asymptotiquement, il soit égal à l'estimateur donné par la minimisation de (17.13) où  $\mathbf{A}(\mathbf{y})$  est indépendant de  $\mathbf{y}$  et égale l'inverse de la matrice de covariance des moments empiriques  $n^{-1/2}\mathbf{F}^\top(\boldsymbol{\theta})\boldsymbol{\iota}$ .

Remarquons que, lorsque la condition nécessaire est vérifiée, la forme de la matrice de covariance asymptotique de l'estimateur GMM  $\hat{\boldsymbol{\theta}}$  se simplifie considérablement. Pour une matrice de pondération limite arbitraire  $\mathbf{A}_0$ , cette matrice était donnée par (17.31). Si la condition est remplie, alors on peut remplacer  $\mathbf{A}_0$  dans (17.31) par l'inverse de  $\boldsymbol{\Phi}$ , qui, selon sa définition (17.29), correspond à la matrice de covariance asymptotique des moments empiriques. Substituant  $\mathbf{A}_0 = \boldsymbol{\Phi}^{-1}$  dans (17.31), nous obtenons le résultat simple selon

lequel

$$V(n^{1/2}(\hat{\theta} - \theta_0)) = (D^\top \Phi^{-1} D)^{-1}.$$

Nous pourrions démontrer le Théorème 17.3 si nous pouvons montrer que, pour toute matrice symétrique, définie positive  $A_0$ , la différence

$$(D^\top A_0 D)^{-1} D^\top A_0 \Phi A_0 D (D^\top A_0 D)^{-1} - (D^\top \Phi^{-1} D)^{-1} \quad (17.37)$$

est semi-définie positive. Pour le montrer, nous récrivons (17.37) sous la forme

$$(D^\top A_0 D)^{-1} D^\top A_0 \left( \Phi - D (D^\top \Phi^{-1} D)^{-1} D^\top \right) A_0 D (D^\top A_0 D)^{-1}. \quad (17.38)$$

Puisque la matrice  $D^\top A_0 D$  est non singulière, (17.38) est définie positive si la matrice que l'on trouve au centre de (17.38), dans le bloc entre parenthèses, l'est. Puisque  $\Phi$  est définie positive, symétrique et de dimension  $l \times l$ , il est possible de trouver une autre matrice définie positive, symétrique et de dimension  $l \times l$  telle que  $\Psi^2 = \Phi^{-1}$ . En termes de  $\Psi$ , la matrice à l'intérieur des parenthèses les plus grandes devient

$$\Psi^{-1} (I - P_{\Psi D}) \Psi^{-1} = \Psi^{-1} M_{\Psi D} \Psi^{-1}, \quad (17.39)$$

où  $P_{\Psi D}$  et  $M_{\Psi D}$  sont, ainsi que le suggèrent les notations, les matrices de projection orthogonale sur l'espace engendré par les colonnes de la matrice  $\Psi D$  de dimension  $l \times k$  et sur son complément orthogonal. Nous voyons que (17.39) est bien une matrice semi-définie positive, ce qui démontre le Théorème 17.3.

Le Théorème 17.3 peut souvent s'interpréter en termes d'**instruments optimaux** ou **poids optimaux**, parce que les conditions du premier ordre pour un minimum de la fonction critère construite avec une matrice de pondération optimale ressemblent fort aux conditions sur les moments empiriques. S'il faut estimer  $k$  paramètres, il y aura précisément  $k$  conditions du premier ordre. Ainsi un modèle qui était à l'origine suridentifié peut être rendu comparable à un modèle juste identifié. Considérons la fonction critère asymptotique  $m^\top(\theta) \Phi^{-1} m(\theta)$  construite à l'aide de la matrice de pondération asymptotique optimale  $\Phi^{-1}$ . Les conditions du premier ordre pour un minimum sont données par les  $k$  composantes de l'équation

$$D^\top(\theta) \Phi^{-1} m(\theta) = 0. \quad (17.40)$$

Supposons que l'on puisse trouver un estimateur convergent  $\hat{\Phi}$  tel que

$$\text{plim}_{\mu} \hat{\Phi} = \Phi(\mu).$$

Si  $D_t(\mathbf{y}, \theta)$  désigne la matrice de dimension  $l \times k$  dont l'élément type est  $\partial f_{ti}(y_t, \theta) / \partial \theta_j$ , (17.23) implique que

$$\text{plim}_{\mu} \left( \frac{1}{n} \sum_{t=1}^n D_t(\mathbf{y}, \theta) \right) = D(\theta).$$

Par conséquent, à l'aide de ces deux équations et de (17.15), la contrepartie empirique à (17.40) est

$$\left( \frac{1}{n} \sum_{t=1}^n \mathbf{D}_t^\top(\mathbf{y}, \boldsymbol{\theta}) \right) \hat{\boldsymbol{\Phi}}^{-1} \left( \frac{1}{n} \sum_{t=1}^n \mathbf{F}_t(\mathbf{y}, \boldsymbol{\theta}) \right). \quad (17.41)$$

Les moments empiriques (17.41) constituent un ensemble de  $k$  combinaisons linéaires des moments d'origine  $n^{-1} \sum_{t=1}^n \mathbf{F}_t$ . En annulant ces équations, nous obtenons  $k$  équations à  $k$  inconnues, et la solution à ces équations est précisément l'estimateur GMM obtenu en minimisant la forme quadratique des moments empiriques élaborée à l'aide d'une matrice de pondération optimale. On peut donner le nom de **moments optimaux** associés à l'ensemble d'origine aux moments (17.41). A l'aide de quelques exemples, nous verrons comment ces moments optimaux peuvent dans bien des cas servir à définir les instruments ou les poids optimaux.

Considérons tout d'abord le cas de l'estimateur IV lorsqu'il y a plus d'instruments que de régresseurs. Les conditions du premier ordre pour la minimisation de la fonction critère (17.08) sont

$$\mathbf{X}^\top \mathbf{P}_W (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}. \quad (17.42)$$

Leur résolution conduit à l'estimateur IV (ou estimateur 2SLS)

$$\tilde{\boldsymbol{\beta}} \equiv (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_W \mathbf{y}, \quad (17.43)$$

qui est identique à l'estimateur IV *simple* obtenu à l'aide des variables instrumentales  $\mathbf{P}_W \mathbf{X}$ . Ainsi l'utilisation optimale de la matrice complète des  $l$  instruments  $\mathbf{W}$  équivaut à l'utilisation des  $k$  instruments que sont les colonnes de la matrice  $\mathbf{P}_W \mathbf{X}$ .

L'estimateur IV en présence d'une hétéroscédasticité de forme inconnue fournit un exemple encore plus intéressant. Dans la section précédente, nous montrions comment construire une HCCME pour l'estimateur IV (17.43) basée sur (17.36). En présence d'hétéroscédasticité cependant, l'estimateur (17.03) ne satisfait plus du tout la condition nécessaire pour l'efficacité asymptotique. Il est possible de construire un estimateur qui satisfait pleinement cette condition en partant des conditions sur les moments (17.05). Soit  $\boldsymbol{\Omega}$  une matrice diagonale de dimension  $n \times n$  dont l'élément type est  $\Omega_{tt} = E(u_t^2)$ , où  $u_t = y_t - \mathbf{X}_t \boldsymbol{\beta}$ . Alors la matrice de covariance des moments empiriques dans (17.05) est simplement  $\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W}$ . Ainsi une fonction critère qui satisfait la condition nécessaire à l'efficacité est

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Les conditions du premier ordre pour un minimum de cette fonction sont

$$\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0},$$

et elles conduisent à l'estimateur

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{y}. \quad (17.44)$$

Les instruments optimaux qui produisent cet estimateur sont les colonnes de la matrice  $\mathbf{W} (\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X}$ . Nous avons ici supposé implicitement que  $\boldsymbol{\Omega}$  est connue. Dans le cas plus réaliste où elle est inconnue, nous pouvons estimer  $\mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W}$  de manière convergente de plusieurs façons, par l'usage des estimateurs non convergents de  $\boldsymbol{\Omega}$  dont nous avons discuté dans la Section 16.3.

Les versions opérationnelles de l'estimateur (17.44) furent proposées à l'origine par Cragg (1983), dans le cas où les régresseurs  $\mathbf{X}$  peuvent être traités comme instruments, et par Cumby, Huizinga, et Obstfeld (1983) dans un cas plus général. Ces derniers considèrent en réalité un estimateur plus compliqué qui permettrait de gérer autant l'hétéroscédasticité que l'autocorrélation, et l'appellèrent estimateur des doubles moindres carrés en deux étapes; nous discuterons de cet estimateur dans la Section 17.5. Nous nous référerons à (17.44) avec  $\boldsymbol{\Omega}$  remplacée par une matrice diagonale de dimension  $n \times n$  dont les éléments diagonaux sont les carrés des résidus 2SLS sous le nom de **H2SLS**, parce qu'il s'agit d'une version modifiée de l'estimateur 2SLS conventionnel qui atteint une efficacité supérieure en présence d'une hétéroscédasticité de forme inconnue. Pareillement, nous appellerons l'estimateur de Cragg, qui emploie les résidus OLS pour estimer  $\boldsymbol{\Omega}$ , estimateur **HOLS**.

Il est révélateur d'examiner plus attentivement ces estimateurs. Si les seuls instruments disponibles sont les régresseurs, alors remplacer  $\mathbf{W}$  par  $\mathbf{X}$  dans (17.44) n'apporte rien de plus et l'on retrouve l'estimateur des OLS. Cragg suggère alors d'employer des puissances ou des produits croisés des régresseurs en tant qu'instruments supplémentaires. Si tous les régresseurs ne peuvent pas servir en tant qu'instruments pour que le modèle soit juste identifié, alors  $\mathbf{W}^\top \mathbf{X}$  est une matrice carrée non singulière et (17.44) se réduit à l'estimateur IV simple. Dans les deux cas, bien évidemment, (17.44) peut ne pas être efficace. Cela nous permet de constater que la condition nécessaire d'efficacité donnée par le Théorème 17.3 *n'est pas* suffisante.

Dans le contexte suridentifié, l'estimateur HOLS sera plus efficace que l'estimateur OLS, et l'estimateur H2SLS sera plus efficace que l'estimateur IV usuel, mais ni l'un ni l'autre ne sera plus efficace dans l'absolu. On peut trouver une exception à cette remarque, lorsqu'il n'y a pas de phénomène d'hétéroscédasticité et que  $\boldsymbol{\Omega}$  correspond à une matrice identité multipliée par un scalaire. Si l'on pose  $\boldsymbol{\Omega} = \sigma^2 \mathbf{I}$  dans (17.44), on obtient l'estimateur IV ordinaire (17.43). Lorsque (17.44) est calculée à l'aide d'une matrice adéquate quelconque  $\hat{\boldsymbol{\Omega}}$ , l'expression différera numériquement de (17.43) lorsque les aléas sont homoscedastiques bien que cette différence s'estompe asymptotiquement. Lorsqu'il y a hétéroscédasticité, nous voyons que si les régresseurs peuvent être traités en tant qu'instruments, l'existence d'*autres* instruments valides peut mener à une amélioration de l'efficacité. Même si tous les

régresseurs ne peuvent pas être utilisés comme instruments, il est possible d'obtenir un gain d'efficacité en utilisant (17.44) au lieu de (17.43). Nous examinerons plus tard la source de ce gain d'efficacité, au cours de la section suivante, lorsque nous considérerons les conditions portant sur les moments conditionnels.

Il nous faut faire quelques remarques à propos des cas où les estimateurs GMM ne sont pas efficaces même si l'on utilise une matrice de pondération optimale. Il s'avère que l'efficacité ou la non efficacité de l'estimateur GMM dépend du modèle sous-jacent  $\mathbb{M}$  pour lequel il est employé. Tout en restant assez vagues, disons que l'estimateur GMM est d'autant plus efficace que le modèle  $\mathbb{M}$  est contraignant. Autrement dit, la probabilité de trouver un estimateur plus efficace que l'estimateur GMM est d'autant plus forte que l'on impose un grand nombre de contraintes dans la spécification de  $\mathbb{M}$ .

Un exemple peut aider à la compréhension de ce point de l'exposé. Considérons un modèle paramétrisé  $(\mathbb{M}_1, \boldsymbol{\theta})$  que l'on peut estimer par maximum de vraisemblance, avec une application définissant des paramètres bi-univoque  $\boldsymbol{\theta} : \mathbb{M}_1 \rightarrow \Theta \subseteq \mathbb{R}^k$ . L'estimateur ML peut être traité comme un estimateur GMM pour lequel les moments empiriques sont les composantes du vecteur score  $\mathbf{g}(\boldsymbol{\theta})$ . L'efficacité asymptotique de l'estimateur du maximum de vraisemblance implique par conséquent celle de l'estimateur GMM. Supposons maintenant que  $\boldsymbol{\theta}$  soit contraint à satisfaire l'égalité vectorielle  $\boldsymbol{\theta}_2 = \mathbf{0}$ , où  $\boldsymbol{\theta}_2$  est un sous-vecteur de dimension  $r$  de  $\boldsymbol{\theta}$ . Ces contraintes définissent un nouveau modèle, contraint, que l'on peut noter  $\mathbb{M}_0$ , tel que  $\mathbb{M}_0 \subset \mathbb{M}_1$ . Grâce au maximum de vraisemblance, le modèle contraint  $\mathbb{M}_0$  peut être estimé exactement de la même manière que le modèle non contraint  $\mathbb{M}_1$ , et l'estimateur ML du premier est en général plus efficace que l'estimateur ML du second.

Dans la structure GMM, les choses peuvent s'exprimer de manière assez différente. Les  $k$  composantes du vecteur score  $\mathbf{g}(\boldsymbol{\theta})$  fournissent  $k$  conditions sur les moments qui devraient être satisfaites par tout DGP de  $\mathbb{M}_1$ , et en particulier par ceux compris dans  $\mathbb{M}_0$ . Si l'on trouve des motivations dans le choix de  $\mathbb{M}_0$ , alors il faudrait sans doute évaluer ces conditions sur les moments en posant le sous-vecteur  $\boldsymbol{\theta}_2$  égal à zéro, mais même ainsi on dispose de  $k$  conditions pour seulement  $k - r$  paramètres; autrement dit, il y a des contraintes de suridentification. La procédure ML les ignore tout simplement et sélectionne juste  $k - r$  de ces conditions, et plus précisément celles fournies par les dérivées partielles de la fonction de logvraisemblance par rapport à  $\boldsymbol{\theta}_1$ . La théorie de l'estimation par maximum de vraisemblance nous enseigne que ce choix est asymptotiquement efficace, et par conséquent, si ces conditions étaient précisément utilisées dans une procédure GMM juste identifiée, celle-ci serait également efficace.

Malgré tout, la procédure GMM usuelle consisterait à construire une forme quadratique à partir de toutes les composantes du gradient et d'une estimation de sa matrice de covariance, qui pourrait être n'importe quelle estimation adéquate de la matrice d'information. Notons  $\hat{J}$  cette estimation,

et nous obtenons

$$\mathbf{g}^\top(\boldsymbol{\theta}_1, \mathbf{0}) \hat{\mathbf{J}}^{-1} \mathbf{g}(\boldsymbol{\theta}_1, \mathbf{0}). \quad (17.45)$$

La minimisation de cette expression par rapport à  $\boldsymbol{\theta}_1$  conduira, en général, à un ensemble d'estimations différent de celui produit par la maximisation de la fonction de logvraisemblance contrainte, mais on peut voir que les deux ensembles sont asymptotiquement équivalents (Cela serait un bon exercice que de le montrer). Cela signifie que l'estimateur GMM est asymptotiquement efficace *à condition* que les contraintes de suridentification soient utilisées.

Les paramètres  $\boldsymbol{\theta}$  peuvent être identifiés dans de nombreux cas par d'autres ensembles de  $k$  conditions portant sur les moments que celles fournies par les dérivées de la fonction de logvraisemblance par rapport à  $\boldsymbol{\theta}_1$ . De façon générale, on peut sélectionner n'importe quel ensemble de  $k - r$  conditions et les résoudre pour obtenir des estimations GMM différentes, qui ne seront pas asymptotiquement efficaces. (Le montrer serait un bon exercice) Il est même envisageable de sélectionner un nombre de conditions compris entre  $k - r$  et  $k$ , de construire une forme quadratique grâce à l'inverse de la matrice d'information, et de minimiser cette forme quadratique afin d'obtenir encore un autre ensemble d'estimations GMM non efficaces.

La conclusion que l'on peut tirer de tout ceci est qu'il existe de multiples possibilités pour un ensemble de conditions sur les moments d'identifier les paramètres d'un modèle  $M_0$ , avec ou sans contrainte de suridentification. Seul un petit nombre de possibilités conduit à des estimations asymptotiquement efficaces. Une discussion détaillée de ces conséquences nous conduirait beaucoup trop loin. Bien qu'il n'existe pas d'obstacle majeur à la compréhension du phénomène dans le contexte ML, un traitement rigoureux dans le cas plus général semble manquer, bien qu'un nombre de cas particuliers soient bien compris. Les lecteurs intéressés peuvent consulter Chamberlain (1986, 1987), Hansen (1985), et Hansen, Heaton, et Ogaki (1988). Heureusement, les choses sont plus simples dans le cas des modèles définis par des conditions portant sur les moments conditionnels, dont nous allons parler dans la prochaine section.

## 17.4 ESTIMATION À L'AIDE DES MOMENTS CONDITIONNELS

Les conditions portant sur les moments employées jusqu'à présent étaient toutes non conditionnelles. Dans la pratique cependant, le fait qu'un modèle économétrique soit spécifié uniquement en termes de moments non conditionnels est l'exception plutôt que la règle. Dans la littérature consacrée aux modèles d'anticipations rationnelles par exemple, la théorie économique requiert que les erreurs de prévision commises par les agents soient indépendantes de toutes les variables de leurs ensembles d'informations à l'instant où les prévisions sont établies. Dans le contexte simple du modèle de régression linéaire  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , il est habituel de supposer non seulement que l'aléa  $u_t$

est non corrélé aux régresseurs  $\mathbf{X}$  mais aussi que son espérance *conditionnelle* aux régresseurs est nulle, ce qui implique à nouveau qu'il est non corrélé avec une fonction quelconque des régresseurs. Dans un contexte de données temporelles, il est très fréquent de supposer que l'erreur  $u_t$  a une espérance nulle conditionnellement à toutes les valeurs passées des régresseurs aussi bien qu'à leurs valeurs courantes.

De façon formelle, il est aisé d'écrire un ensemble d'équations définissantes des paramètres en termes des moments conditionnels. Il n'y a souvent qu'une seule équation de ce genre, que l'on peut écrire

$$E(f_t(y_t, \boldsymbol{\theta}) | \Omega_t) = 0 \quad \text{pour tout } t = 1, \dots, n, \quad (17.46)$$

où  $\Omega_t$  est l'ensemble d'informations pour l'observation  $t$ . Nous ferons l'hypothèse simplificatrice que  $\Omega_t \subseteq \Omega_s$  pour  $t < s$ . Dans (17.46) nous interprétons  $f_t(y_t, \boldsymbol{\theta})$  comme une sorte d'erreur, telle qu'une erreur de prévision commise par les agents économiques. Le cas d'une estimation IV d'un modèle de régression linéaire offre un exemple simple. Dans ce cas précis, (17.46) nous indique que les erreurs, une par observation, sont orthogonales à l'ensemble d'informations défini par l'ensemble des instruments. Il serait possible d'avoir plusieurs équations définissantes des paramètres telles que (17.46), comme dans le cas d'un modèle de régression multivariée, mais pour simplifier nous supposerons dans cette section qu'il n'en existe qu'une seule.

En théorie, aucun problème d'identification ne se pose du fait qu'il n'existe qu'une seule équation définissante des paramètres, parce qu'il existe un nombre infini d'instruments possibles dans le genre d'ensemble d'informations que nous considérons. Dans la pratique, bien évidemment, il faut choisir un nombre fini d'instruments, afin d'établir une fonction critère pour l'estimation GMM. La plus grande partie de cette section consistera à établir les quelques résultats qui affectent ce choix. Nous montrerons que la précision de l'estimateur GMM est reliée positivement au nombre des instruments. Puis, nous montrons que, malgré ce premier résultat, les matrices de covariance asymptotique des estimateurs GMM construits à partir des instruments compris dans les ensembles d'informations  $\Omega_t$  sont bornées inférieurement. La borne inférieure, qui s'apparente à la borne inférieure de Cramér-Rao introduite dans le Chapitre 8, est souvent appelée **borne GMM**. En théorie, tout au moins, il existe un ensemble optimal d'instruments qui permet d'atteindre la borne GMM, et les instruments optimaux peuvent dans certains cas être calculés ou estimés.

Nous construisons un ensemble de  $l$  instruments  $\mathbf{w}_1, \dots, \mathbf{w}_l$  que l'on peut grouper dans une matrice  $\mathbf{W}$  de dimension  $n \times l$  telle que  $W_{ti} \in \Omega_t$  pour tout  $t = 1, \dots, n$  et  $i = 1, \dots, l$ . Nous réclamons bien évidemment que  $l \geq k$ , où  $k$  est le nombre de composantes du vecteur de paramètres  $\boldsymbol{\theta}$ . On peut exprimer les conditions portant sur les moments conditionnels que l'on utilise pour l'estimation comme suit:

$$\mathbf{W}^\top \mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}, \quad (17.47)$$

où  $\mathbf{f}$  est un vecteur à  $n$  composantes, et dont la composante type est  $f_t$ . Si  $l = k$ , l'estimateur  $\hat{\boldsymbol{\theta}}$  est obtenu en résolvant les  $k$  équations (17.47). Si  $l > k$ , cet estimateur est obtenu en minimisant la forme quadratique élaborée à partir des composantes du membre de gauche de (17.47) et d'une estimation de leur matrice de covariance. Notons  $\boldsymbol{\Omega}$  la matrice de covariance des  $f_t$ . Ainsi, si nous notons  $\mu$  le DGP et  $\boldsymbol{\theta}_0$  le véritable vecteur de paramètres,

$$\Omega_{ts} = E_{\mu}(f_t(\boldsymbol{\theta}_0)f_s(\boldsymbol{\theta}_0) | \Omega_t) \quad \text{pour tout } t \leq s.$$

Alors la matrice de covariance conditionnelle des moments empiriques dans (17.47) est  $\boldsymbol{\Phi} \equiv \mathbf{W}^{\top} \boldsymbol{\Omega} \mathbf{W}$ .

Dans le cas habituel, où  $l > k$ , la fonction critère utilisée pour obtenir les estimations des paramètres est

$$\mathbf{f}(\boldsymbol{\theta})^{\top} \mathbf{W} (\mathbf{W}^{\top} \boldsymbol{\Omega} \mathbf{W})^{-1} \mathbf{W}^{\top} \mathbf{f}(\boldsymbol{\theta}).$$

La matrice de covariance asymptotique de cet estimateur est donnée par la limite en probabilité de  $(\mathbf{D}^{\top} \boldsymbol{\Phi}^{-1} \mathbf{D})^{-1}$ , où

$$\mathbf{D}_{ij} = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n W_{ti} \frac{\partial f_t}{\partial \theta_j} \right). \quad (17.48)$$

Soit  $\mathbf{J}(\mathbf{y}, \boldsymbol{\theta})$  la matrice de dimension  $n \times k$  d'élément type  $\partial f_t(y_t, \boldsymbol{\theta}) / \partial \theta_j$ .<sup>1</sup> Alors le membre de droite de (17.48) est la limite de  $n^{-1} \mathbf{W}^{\top} \mathbf{J}$ . Par conséquent, la matrice de covariance asymptotique de  $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  se résume à la limite de

$$\left( \left( \frac{1}{n} \mathbf{J}^{\top} \mathbf{W} \right) \left( \frac{1}{n} \mathbf{W}^{\top} \boldsymbol{\Omega} \mathbf{W} \right)^{-1} \left( \frac{1}{n} \mathbf{W}^{\top} \mathbf{J} \right) \right)^{-1}. \quad (17.49)$$

Le premier résultat relatif au choix optimal des instruments  $\mathbf{W}$  est simple et intuitif. Il indique que si nous augmentons le nombre des instruments, la matrice de covariance limite (17.49) ne peut pas augmenter. Imaginons qu'au lieu des conditions portant sur les moments empiriques (17.47) nous utilisons un ensemble de combinaisons linéaires de ces conditions. Cela correspond à

$$\mathbf{B}^{\top} \mathbf{W}^{\top} \mathbf{f}(\boldsymbol{\theta}) = \mathbf{0},$$

à la place de (17.47), pour une matrice  $\mathbf{B}$  de dimension  $l \times p$  quelconque, où  $p \leq l$ . Il est aisé de voir que cela correspond au remplacement de  $\mathbf{D}$  par  $\mathbf{B}^{\top} \mathbf{D}$  et de  $\boldsymbol{\Phi}$  par  $\mathbf{B}^{\top} \boldsymbol{\Phi} \mathbf{B}$ . Considérons la différence

$$\mathbf{D}^{\top} \boldsymbol{\Phi}^{-1} \mathbf{D} - \mathbf{D}^{\top} \mathbf{B} (\mathbf{B}^{\top} \boldsymbol{\Phi} \mathbf{B})^{-1} \mathbf{B}^{\top} \mathbf{D}$$

<sup>1</sup> La notation  $\mathbf{J}$  fut choisie parce que la matrice est la matrice *Jacobienne* de  $\mathbf{f}$  par rapport à  $\boldsymbol{\theta}$  et parce que  $\mathbf{F}$  était déjà réservée à un autre usage.



entre les inverses des matrices de covariance asymptotique de dimension  $k \times k$  correspondant aux instruments  $\mathbf{W}$  et  $\mathbf{WB}$ , respectivement. Si, comme précédemment, nous notons  $\Psi$  une matrice symétrique de dimension  $l \times l$  telle que  $\Psi^2 = \Phi^{-1}$ , cette différence devient

$$D^\top \Psi \left( \mathbf{I} - \Psi^{-1} \mathbf{B} (\mathbf{B}^\top \Psi^{-2} \mathbf{B})^{-1} \mathbf{B}^\top \Psi^{-1} \right) \Psi D. \quad (17.50)$$

Cette matrice est à l'évidence semi-définie positive, parce que la matrice entre les deux grandes parenthèses est la matrice de projection orthogonale sur le complément orthogonal de l'espace engendré par les colonnes de  $\Psi^{-1} \mathbf{B}$ . Pour deux matrices quelconques  $\mathbf{P}$  et  $\mathbf{Q}$ , symétriques, définies positives et de même dimension,  $\mathbf{P} - \mathbf{Q}$  est semi-définie positive si et seulement si  $\mathbf{Q}^{-1} - \mathbf{P}^{-1}$  est semi-définie positive (consulter l'Annexe A). Ainsi le fait que (17.50) soit semi-définie positive établit notre premier résultat.

Ce résultat semble suggérer qu'il faudrait utiliser autant d'instruments que possible afin d'obtenir des estimations aussi efficaces que possible. Malgré tout, une telle conclusion est généralement fausse. Souvenons-nous de la discussion de la Section 7.5, illustrée par la Figure 7.1. Nous avons vu que, dans le contexte IV ordinaire, il y a un équilibre à réaliser entre l'efficacité asymptotique et le biais avec des échantillons finis. Le même équilibre doit également être recherché dans le cas GMM. L'usage d'un nombre important de contraintes de suridentification peut mener à une matrice de covariance asymptotique plus petite, mais les estimations peuvent se révéler très sévèrement biaisées. Un autre argument allant à l'encontre de l'usage d'un trop grand nombre d'instruments est simplement que les conséquences positives sont décroissantes, compte tenu de l'existence de la borne GMM.

Le second résultat montre comment choisir les instruments  $\mathbf{W}$  de façon optimale. Il indique que si nous posons  $\mathbf{W} = \Omega^{-1} \mathbf{J}$  dans (17.47), la matrice de covariance asymptotique qui en résulte est plus petite que celle donnée par n'importe quel autre choix. A partir de (17.49) il s'ensuit que la borne GMM pour la matrice de covariance asymptotique est  $\text{plim } (n^{-1} \mathbf{J}^\top \Omega^{-1} \mathbf{J})^{-1}$ . Hélas, comme nous le verrons, ce résultat n'est pas toujours opérationnel dans la pratique.

La démonstration est très simple. Comme pour le premier résultat, il est très facile de manipuler des inverses de matrices de covariance pertinentes. Définissons par  $\mathbf{r}$  la matrice symétrique de dimension  $n \times n$  telle que  $\mathbf{r}^2 \equiv \Omega$ . Alors, la suppression des limites et des facteurs de  $n$  pour l'instant nous montre que

$$\begin{aligned} & \mathbf{J}^\top \Omega^{-1} \mathbf{J} - \mathbf{J}^\top \mathbf{W} (\mathbf{W}^\top \Omega \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{J} \\ &= \mathbf{J}^\top \mathbf{r}^{-1} \left( \mathbf{I} - \mathbf{r} \mathbf{W} (\mathbf{W}^\top \mathbf{r}^2 \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{r} \right) \mathbf{r}^{-1} \mathbf{J}. \end{aligned} \quad (17.51)$$

Puisque la matrice dans les grandes parenthèses est la projection orthogonale sur le complément de l'espace engendré par les colonnes de  $\mathbf{r} \mathbf{W}$ , cette expression est semi-définie positive, et le second résultat est établi.

Il est tout à fait possible que la  $t^{\text{ième}}$  ligne  $\mathbf{J}_t$  de la matrice  $\mathbf{J}$  n'appartienne pas à l'ensemble d'informations  $\Omega_t$ . Dans ce cas, il ne faut surtout pas ignorer les limites et les facteurs de  $n$  dans (17.51). Chaque expression matricielle tend alors vers une limite en probabilité déterministe, qui en vertu de la loi des grands nombres, est la limite des espérances (conditionnelles) des matrices. Par conséquent,  $\mathbf{J}_t$  devrait être remplacée par  $E(\mathbf{J}_t | \Omega_t)$  lorsque cela est nécessaire.

Remarquons que  $\mathbf{\Omega}^{-1}\mathbf{J}$  est une matrice qui possède  $k$  instruments. Nous avons donc montré que, dans le contexte d'un modèle avec des conditions portant sur les moments conditionnels, il est possible de choisir des instruments tels que, bien qu'il n'y ait aucune contrainte de suridentification, on obtienne un estimateur asymptotiquement efficace. La matrice de covariance asymptotique associée à cet estimateur est  $\text{plim}(n^{-1}\mathbf{J}^\top\mathbf{\Omega}^{-1}\mathbf{J})$ . Dans la pratique, il peut être plus ou moins facile de calculer ou d'estimer les instruments optimaux. Clairement, la matrice  $\mathbf{J}(\boldsymbol{\theta})$  peut se calculer directement comme une fonction de  $\boldsymbol{\theta}$  en dérivant les moments empiriques. Mais il faut ensuite une estimation de  $\boldsymbol{\theta}$ , à moins que les moments ne soient linéaires par rapport à  $\boldsymbol{\theta}$ . Une attitude à adopter consiste à obtenir en premier lieu une estimation convergente mais non efficace et de l'utiliser pour définir de façon approximative les instruments optimaux, qui nous conduiront ensuite à des estimations asymptotiquement efficaces. Si les estimations de départ ne sont pas très précises, il serait grandement souhaitable d'employer une procédure itérative au cours de laquelle des estimations successives définissent des approximations successives de plus en plus proches des instruments optimaux.

Afin d'obtenir des instruments optimaux, il est également nécessaire d'estimer la matrice  $\mathbf{\Omega}$  de façon convergente, au moins à un facteur multiplicatif près. Si les  $f_t$  sont homoscédastiques et indépendants en série, on peut bien sûr employer simplement une matrice identité pour  $\mathbf{\Omega}$ . Si elles suivent une structure connue d'hétéroscédasticité et/ou d'autocorrélation, avec des paramètres que l'on peut estimer de façon convergente, alors il est envisageable d'employer une procédure itérative ou une procédure en deux étapes. Mais s'il peut y avoir une structure d'hétéroscédasticité ou d'autocorrélation arbitraire, cela devient un sujet, sinon désespéré, du moins extrêmement délicat à traiter. Habituellement, les instruments optimaux ne peuvent plus être calculés et il faut se contenter des instruments disponibles.

Voyons à présent comment appliquer les résultats de cette section à un cas simple. Considérons le modèle de régression linéaire pour lequel les ensembles d'informations  $\Omega_t$  sont connus pour chaque observation. La condition sur le moment qui définit le vecteur de paramètres  $\boldsymbol{\beta}$  est  $E(y_t - \mathbf{X}_t\boldsymbol{\beta} | \Omega_t) = 0$ . En termes de notre notation générale,  $f_t = y_t - \mathbf{X}_t\boldsymbol{\beta}$ , et la matrice  $\mathbf{J}$  est simplement égale à  $\mathbf{X}$ . De façon comparable, la matrice  $\mathbf{\Omega}$  correspond simplement à la matrice de covariance des  $f_t$ , c'est-à-dire celle des aléas. Ainsi, à condition que  $\mathbf{X}_t \in \Omega_t$ , les instruments optimaux sont donnés par les colonnes

de  $\Omega^{-1}\mathbf{X}$ . Les conditions portant sur les moments empiriques deviennent

$$\mathbf{X}^\top \Omega^{-1}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0},$$

et nous voyons que, comme nous aurions pu nous y attendre, l'estimateur efficace est celui des GLS.

Cet exemple devrait montrer au moins certains aspects des difficultés qui peuvent entacher le calcul des instruments optimaux. Comme nous l'avons vu dans la Section 9.5, si la forme de la matrice  $\Omega$  est connue et dépend d'un vecteur de paramètres que l'on peut estimer de façon convergente à partir d'une procédure auxiliaire, les GLS faisables produisent des estimations asymptotiquement équivalentes à celles d'une véritable procédure GLS. De façon similaire, dans un contexte de GMM, si la forme de  $\Omega$  est connue, il est envisageable d'estimer les instruments optimaux et d'obtenir des estimations GMM asymptotiquement efficaces. Cependant, il n'est pas rare que  $\Omega$  soit inconnue et ne puisse pas être estimée de façon convergente. Nous verrons comment gérer de telles circonstances dans la section qui suit.

Il est relativement aisé d'étendre la procédure des GLS discutée plus haut au cas où certains éléments de  $\mathbf{X}_t$  n'appartiennent pas à l'ensemble  $\Omega_t$  et où des variables instrumentales doivent être utilisées. Comme nous l'avons vu,  $\mathbf{J}_t$  doit être remplacée dans ce cas par son espérance conditionnelle à  $\Omega_t$  dans la définition des instruments optimaux, qui correspondent alors aux colonnes de  $\Omega^{-1}E(\mathbf{X}_t | \Omega_t)$ . Dans le cas particulier d'erreurs homoscédastiques et non autocorrélées, ce résultat nous apprend que les meilleures variables instrumentales à utiliser sont les espérances des régresseurs conditionnellement à toutes les variables qui sont orthogonales aux aléas. Dans la pratique, ces espérances conditionnelles peuvent ne pas être disponibles, et il faut alors se contenter des instruments dont on dispose.

Si  $\Omega$  est connue ou peut être estimée par une procédure faisable, on peut choisir un ensemble disponible d'instruments  $\mathbf{W}$  et former les conditions sur les moments empiriques

$$\mathbf{W}^\top \Omega^{-1}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}. \quad (17.52)$$

Il devrait normalement y avoir plus d'instruments que de paramètres, puisque les instruments optimaux ne sont pas disponibles et que les contraintes de suridentification amélioreront par conséquent l'efficacité. Afin de satisfaire la condition nécessaire du Théorème 17.3, la fonction critère doit utiliser la matrice de covariance du membre de gauche de (17.52). Celle-ci est, asymptotiquement,

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{W}^\top \Omega^{-1}(\mathbf{y} - \mathbf{X}\beta)(\mathbf{y} - \mathbf{X}\beta)^\top \Omega^{-1} \mathbf{W} \right) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{W}^\top \Omega^{-1} \mathbf{W} \right).$$

La fonction critère pertinente est par conséquent

$$(\mathbf{y} - \mathbf{X}\beta)^\top \Omega^{-1} \mathbf{W} (\mathbf{W}^\top \Omega^{-1} \mathbf{W})^{-1} \mathbf{W}^\top \Omega^{-1} (\mathbf{y} - \mathbf{X}\beta),$$

qui conduit aux conditions du premier ordre

$$\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{W} (\mathbf{W}^\top \boldsymbol{\Omega}^{-1} \mathbf{W})^{-1} \mathbf{W}^\top \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}. \quad (17.53)$$

Cette équation définit un estimateur apparemment bien compliqué. En vérité, on peut l'interpréter assez simplement, tout comme l'estimateur GLS, en termes d'une matrice de transformation  $\boldsymbol{\eta}$  telle que  $\boldsymbol{\eta}^\top \boldsymbol{\eta} = \boldsymbol{\Omega}^{-1}$ . Soit

$$\mathbf{y}^* \equiv \boldsymbol{\eta} \mathbf{y}, \quad \mathbf{X}^* \equiv \boldsymbol{\eta} \mathbf{X}, \quad \text{et} \quad \mathbf{Z} \equiv \boldsymbol{\eta} \mathbf{W}.$$

Alors (17.53) devient

$$\mathbf{X}^{*\top} \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top (\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}) = \mathbf{X}^{*\top} \mathbf{P}_Z (\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}) = \mathbf{0}.$$

Cette équation définit un estimateur IV ordinaire en termes des variables transformées  $\mathbf{y}^*$  et  $\mathbf{X}^*$  et des transformations des instruments  $\mathbf{Z}$ . Ainsi, l'estimateur défini par (17.53) peut être calculé sans plus de difficulté que l'estimateur GLS. Cet estimateur est pertinent chaque fois que les GLS ou les GLS faisables auraient été appropriés sauf s'il y a une éventuelle corrélation entre les aléas et les régresseurs.

L'estimateur défini par (17.53) porte en lui une lourde ressemblance avec l'estimateur H2SLS (17.44) défini dans la section précédente. En réalité, la substitution de  $\boldsymbol{\Omega}^{-1} \mathbf{W}$  à  $\mathbf{W}$  permet de passer du premier au second. La théorie développée dans cette section montre que s'il est possible de choisir  $\mathbf{W}$  comme les espérances conditionnelles des régresseurs  $\mathbf{X}$  (ou des combinaisons linéaires de ceux-ci), alors l'estimateur défini par (17.53) est asymptotiquement efficace, et l'estimateur H2SLS ne l'est pas. L'avantage de l'estimateur H2SLS est qu'il peut être calculé en présence d'une hétéroscédasticité dont la forme est inconnue, puisque  $n^{-1} \mathbf{W}^\top \boldsymbol{\Omega} \mathbf{W}$  peut être estimée de façon convergente en employant des estimateurs non convergents de  $\boldsymbol{\Omega}$ . Par contre, (17.53) ne peut être formulé qu'à condition que  $\boldsymbol{\Omega}$  soit elle-même estimée de façon convergente, parce que des expressions telles que  $n^{-1} \mathbf{W}^\top \boldsymbol{\Omega}^{-1} \mathbf{W}$  et  $n^{-1} \mathbf{W}^\top \boldsymbol{\Omega}^{-1} \mathbf{y}$  ne peuvent pas être estimées de façon convergente sans une estimation elle-même convergente de  $\boldsymbol{\Omega}$ . Ainsi les deux estimateurs se révèlent utiles, mais dans des circonstances différentes.

Le concept de borne GMM fut introduit, non pas sous ce nom, par Hansen (1985), qui donna également les conditions pour les instruments optimaux. Cependant, les arguments utilisés pour dériver la borne ont une longue histoire, et Hansen date la recherche des instruments efficaces à Basman (1957) et Sargan (1958).

## 17.5 ESTIMATION DE LA MATRICE DE COVARIANCE

Dans les sections précédentes, nous avons fait allusion aux difficultés que l'on peut rencontrer lors de l'estimation des matrices de covariance dans le contexte de la GMM. En vérité, les problèmes surviennent de deux sources différentes: la première pour le choix de la matrice de pondération à utiliser lors de la construction de la fonction critère et la seconde pour l'estimation proprement dite de la matrice de covariance des estimations. Par chance, des considérations semblables s'appliquent aux deux problèmes, de sorte que l'on peut les traiter simultanément.

Souvenons-nous à partir de (17.31) que la matrice de covariance asymptotique d'un estimateur GMM calculé à l'aide de la matrice de pondération  $\mathbf{A}_0$  est

$$(\mathbf{D}^\top \mathbf{A}_0 \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{A}_0 \boldsymbol{\Phi} \mathbf{A}_0 \mathbf{D} (\mathbf{D}^\top \mathbf{A}_0 \mathbf{D})^{-1},$$

en conservant la notation de la Section 17.2. Si la condition nécessaire à l'efficacité donnée par le Théorème 17.3 est satisfaite, on doit avoir  $\mathbf{A}_0 \stackrel{a}{=} \boldsymbol{\Phi}^{-1}$ , où  $\boldsymbol{\Phi}$  est la matrice de covariance asymptotique de dimension  $l \times l$  des moments empiriques  $n^{-1/2} \mathbf{F}^\top(\boldsymbol{\theta}) \boldsymbol{\iota}$  dont l'élément type est

$$n^{-1/2} \sum_{t=1}^n f_{ti}(y_t, \boldsymbol{\theta}).$$

Ainsi le problème consiste à trouver un estimateur convergent  $\hat{\boldsymbol{\Phi}}$  de  $\boldsymbol{\Phi}$ . Si cela est possible, alors nous pouvons minimiser la fonction critère

$$\boldsymbol{\iota}^\top \mathbf{F}(\boldsymbol{\theta}) \hat{\boldsymbol{\Phi}}^{-1} \mathbf{F}^\top(\boldsymbol{\theta}) \boldsymbol{\iota}. \quad (17.54)$$

Si un élément type de  $\hat{\mathbf{D}}$  est défini par (17.32), la matrice de covariance asymptotique de  $\hat{\boldsymbol{\theta}}$  peut être estimée par

$$\frac{1}{n} (\hat{\mathbf{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{D}})^{-1}. \quad (17.55)$$

Il est clair qu'il nous faut procéder en au moins deux étapes, parce que  $\hat{\boldsymbol{\Phi}}$  doit être une estimation de la matrice de covariance des moments empiriques *évaluée avec les véritables valeurs des paramètres*. Ainsi avant que  $\hat{\boldsymbol{\Phi}}$  ne puisse être calculée, il est nécessaire de disposer au préalable d'un estimateur convergent des paramètres  $\boldsymbol{\theta}$ . Puisque l'on peut employer une matrice de pondération  $\mathbf{A}_0$  arbitraire sans perte de convergence, il y a plusieurs façon d'obtenir cette estimation préliminaire. Ensuite,  $\hat{\boldsymbol{\Phi}}$  peut être calculée, et, en minimisant (17.54), fournir un nouvel ensemble d'estimations des paramètres. Il est possible de répéter ces opérations successives une ou plusieurs fois si cela s'avère utile. En théorie, une seule itération suffit à obtenir l'efficacité asymptotique mais, dans la pratique, les estimations initiales peuvent se révéler assez mauvaises et cela justifie la multiplication des itérations.

Notre définition précédente de  $\Phi$ , (17.29), se basait sur l'hypothèse que les moments empiriques  $f_{ti}$  étaient indépendants entre eux. Puisque nous souhaitons relâcher cette hypothèse dans cette section, il est nécessaire d'adopter une nouvelle définition de  $\Phi$ , de façon à ce qu'elle reste toujours la matrice de covariance asymptotique des moments empiriques. Nous posons donc la définition:

$$\Phi \equiv \lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n E_{\mu}(\mathbf{F}_t^{\top}(y_t, \theta_0) \mathbf{F}_s(y_t, \theta_0)) \right), \quad (17.56)$$

où  $\mathbf{F}_t$  est la  $t^{\text{ième}}$  ligne de la matrice  $\mathbf{F}$  de dimension  $n \times l$ . Puisque c'est au DGP  $\mu$  que nous faisons référence dans ce qui suit, nous l'enlevons de la notation. L'expression (17.56) diffère de (17.29) en ce qu'elle permet n'importe quel schéma de corrélation entre les contributions  $\mathbf{F}_t$  aux moments empiriques et qu'elle reste valable même si aucun théorème de la limite centrale ne l'est. Il est nécessaire, bien sûr, de supposer que la limite dans (17.56) existe. Notre but est désormais de trouver un estimateur convergent de (17.56).

La première étape consiste à définir les **autocovariances** des moments empiriques

$$\Gamma(j) = \begin{cases} \frac{1}{n} \sum_{t=j+1}^n E(\mathbf{F}_t^{\top}(\theta_0) \mathbf{F}_{t-j}(\theta_0)) & \text{pour } j \geq 0 \\ \frac{1}{n} \sum_{t=-j+1}^n E(\mathbf{F}_{t+j}^{\top}(\theta_0) \mathbf{F}_t(\theta_0)) & \text{pour } j < 0. \end{cases} \quad (17.57)$$

En termes des matrices de dimension  $l \times l$   $\Gamma(j)$ , le membre de droite de (17.56) sans la limite devient

$$\Phi^n \equiv \sum_{j=-n+1}^{n-1} \Gamma(j). \quad (17.58)$$

S'il n'y avait pas de corrélation entre les observations successives, alors seule  $\Gamma(0)$  serait différente de la matrice nulle, et nous aurions

$$\Phi^n = \Gamma(0) = \frac{1}{n} \sum_{t=1}^n E(\mathbf{F}_t^{\top}(\theta_0) \mathbf{F}_t(\theta_0)). \quad (17.59)$$

Puisque le cas de l'indépendance en série est souvent évoqué, il est utile d'examiner deux exemples concrets. Considérons le modèle de régression linéaire  $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ , où  $\mathbf{X}$  est une matrice de dimension  $n \times k$  et où  $\mathbf{W}$  est une matrice d'instruments de dimension  $n \times k$ . Pour ce modèle, qui est juste identifié,

$$\mathbf{F}_t(\beta) = \mathbf{W}_t(y_t - \mathbf{X}_t\beta). \quad (17.60)$$

Ainsi, à partir de (17.59), nous obtenons

$$\Phi^n = \frac{1}{n} \sum_{t=1}^n E(u_t^2) \mathbf{W}_t^\top \mathbf{W}_t, \quad u_t \equiv y_t - \mathbf{X}_t \beta_0. \quad (17.61)$$

Si la véritable matrice de covariance des aléas  $\mathbf{u}$  est la matrice diagonale  $\Omega$ , alors nous avons vu dans la Section 16.3 que nous pouvons estimer  $\lim(n^{-1} \mathbf{W}^\top \Omega \mathbf{W})$  de façon convergente par (17.61) sans l'espérance et en remplaçant  $\beta_0$  par un quelconque estimateur convergent  $\hat{\beta}$ . L'estimateur défini par les moments empiriques (17.60) correspond à l'estimateur IV habituel  $(\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{y}$ , et donc, en utilisant (17.33) et (17.31), nous voyons que sa matrice de covariance asymptotique peut être estimée par

$$\left(\frac{1}{n} \mathbf{W}^\top \mathbf{X}\right)^{-1} \left(\frac{1}{n} \mathbf{W}^\top \hat{\Omega} \mathbf{W}\right) \left(\frac{1}{n} \mathbf{X}^\top \mathbf{W}\right)^{-1}, \quad (17.62)$$

où  $\hat{\Omega}$  est la matrice diagonale de dimension  $n \times n$  dont l'élément type est  $\hat{u}_t^2$ , le carré du  $t^{\text{ième}}$  résidu IV. Cette expression a la forme d'une HCCME standard (voir la Section 16.3). Si le nombre d'instruments dans  $\mathbf{W}$  est supérieur au nombre de régresseurs dans  $\mathbf{X}$ , nous pouvons, tout comme dans (17.43), remplacer simplement  $\mathbf{W}$  par  $\mathbf{P}_W \mathbf{X}$ . Après cette substitution, la limite de (17.62) devient identique à (17.36).

Nous avons noté plus tôt qu'un estimateur de  $\Phi$  peut être utilisé pour deux raisons bien distinctes: estimer la matrice de covariance de n'importe quel ensemble d'estimations GMM et estimer la matrice de pondération optimale. Nous venons juste de fournir un exemple du premier usage, en reconstituant la HCCME dans le cadre d'une estimation par IV. Nous examinons à présent un exemple du second usage, en reconstruisant l'estimateur H2SLS de la Section 17.3. Souvenons-nous que cet estimateur est en général plus efficace que celui des OLS ou des IV en présence d'hétéroscédasticité de forme inconnue.

Les moments empiriques sont les  $l$  composantes de  $\mathbf{W}^\top (\mathbf{y} - \mathbf{X}\beta)$ , où  $l > k$ , et notre estimation de leur matrice de covariance asymptotique est  $\mathbf{W}^\top \hat{\Omega} \mathbf{W}$ . L'inverse de cette estimation peut être employée en tant que matrice de pondération dans la fonction critère

$$(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{W} (\mathbf{W}^\top \hat{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\beta).$$

Les conditions du premier ordre pour un minimum de cette fonction critère sont données par

$$\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \hat{\Omega} \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\beta) = \mathbf{0},$$

et leur résolution conduit à l'estimateur H2SLS (17.44), où l'estimateur  $\hat{\Omega}$  remplace  $\Omega$ .

Il est assez tentant de supposer que, tout comme dans le cas des HCCME, nous pouvons estimer les autocovariances (17.57) simplement en ne calculant pas les espérances dans cette expression, en évaluant les  $\mathbf{F}_t$  avec une estimation préliminaire convergente  $\hat{\boldsymbol{\theta}}$ , et en substituant les  $\hat{\boldsymbol{\Gamma}}(j)$  ainsi obtenus dans (17.58) afin d'aboutir à une estimation adéquate de  $\boldsymbol{\Phi}$ . Hélas, tout n'est pas aussi simple. La **matrice d'autocovariance empirique** à l'ordre zéro,  $\hat{\boldsymbol{\Gamma}}(0)$ , correspond à (17.59) sans l'espérance et évaluée en  $\hat{\boldsymbol{\theta}}$ . Il s'agit d'un estimateur convergent de la véritable matrice d'autocovariance à l'ordre zéro  $\boldsymbol{\Gamma}(0)$ . Mais la matrice de covariance empirique  $\hat{\boldsymbol{\Gamma}}(j)$  à l'ordre  $j$  *ne converge pas* vers la véritable matrice d'autocovariance à l'ordre  $j$  pour un  $j$  arbitraire tel que  $-n+1 \leq j \leq n-1$ . La raison n'est pas difficile à comprendre. Supposons par exemple que  $j = n-2$ . Alors, à partir de (17.57), nous voyons que  $\boldsymbol{\Gamma}(j)$ , et donc aussi  $\hat{\boldsymbol{\Gamma}}(j)$ , ne possède que deux termes. Aucune loi des grands nombres ne peut raisonnablement s'appliquer à deux termes, et  $\hat{\boldsymbol{\Gamma}}(j)$  tend vers zéro lorsque  $n \rightarrow \infty$  à cause du terme  $n^{-1}$  de la définition.

Cette observation suggère un moyen de contourner la difficulté. Nous pourrions par exemple limiter notre attention aux modèles pour lesquels l'autocovariance d'ordre  $j$  *tend effectivement* vers zéro lorsque  $j \rightarrow \infty$ . Si les processus aléatoires qui définissent un DGP possèdent la propriété d'être *mixants* telle que dans la Définition 4.13, nous pouvons montrer que les autocovariances tendent effectivement vers zéro. (Consulter la discussion qui fait suite à la Définition 4.13) Alors il semblerait raisonnable de *tronquer* la somme dans (17.58) en éliminant les termes pour lesquels  $|j|$  est supérieur à une borne choisie.

Si nous notons  $p$  cette borne, nous aurons l'estimateur suivant pour  $\boldsymbol{\Phi}$ :

$$\hat{\boldsymbol{\Phi}} = \hat{\boldsymbol{\Gamma}}(0) + \sum_{j=1}^p (\hat{\boldsymbol{\Gamma}}(j) + \hat{\boldsymbol{\Gamma}}(j)^\top), \quad (17.63)$$

où nous avons utilisé la propriété  $\boldsymbol{\Gamma}(-j) = \boldsymbol{\Gamma}(j)^\top$ , qui provient directement de la définition (17.57). Il est possible de modifier (17.63) en introduisant une correction sur les degrés de liberté sous la forme du facteur  $n/(n-k)$  étant donné que  $k$  paramètres ont été estimés. Mais la pertinence d'une telle procédure avec de petits échantillons mérite d'être encore approfondie.

L'estimateur (17.63) fut proposé par Hansen (1982) et White et Domowitz (1984), et fut employé dans les premières publications qui utilisaient l'estimation par GMM, telles que celle de Hansen et Singleton (1982). D'un point de vue théorique, il est nécessaire de laisser le paramètre de troncature  $p$ , auquel on fait souvent référence en tant que **paramètre de troncature des retards**, diverger à un taux bien précis. Un tel taux serait  $n^{1/4}$ , au quel cas  $p = o(n^{1/4})$ . Cela garantit que, pour un  $n$  suffisamment grand, toutes les  $\boldsymbol{\Gamma}(j)$  non nulles sont estimées de manière convergente. Malheureusement, ce genre de résultat n'est pas transposable dans la pratique, où l'on dispose d'un échantillon de taille  $n$  donnée. Nous reviendrons sur ce point un peu plus tard,



mais nous supposons pour l'instant que nous sommes capables de sélectionner une valeur de  $p$  appropriée.

Une difficulté beaucoup plus sérieuse associée à (17.63) est que, avec des échantillons finis, elle peut très bien ne pas être définie positive ni même semi-définie positive. Si l'on est vraiment malchanceux en disposant d'un ensemble de données qui produit une matrice  $\hat{\Phi}$  non définie, alors (17.63) est inutilisable. Il existe de nombreux moyens de contourner la difficulté. Le plus largement répandu est celui suggéré par Newey et West (1987a). Il consiste simplement à multiplier  $\hat{\Gamma}(j)$  par une série de poids qui décroissent avec  $|j|$ . Typiquement, l'estimateur qu'ils proposent est

$$\hat{\Phi} = \hat{\Gamma}(0) + \sum_{j=1}^p \left(1 - \frac{j}{p+1}\right) (\hat{\Gamma}(j) + \hat{\Gamma}(j)^\top). \quad (17.64)$$

On peut montrer que les poids  $1 - j/(p+1)$  diminuent linéairement avec  $j$  d'une valeur de 1 pour  $\hat{\Gamma}(0)$  par incréments de  $1/(p+1)$  jusqu'à atteindre la valeur  $1/(p+1)$  pour  $|j| = p$ . L'usage de cet ensemble de poids est à l'évidence compatible avec l'idée que l'effet de l'autocovariance d'ordre  $j$  diminue avec  $|j|$ .

Nous n'essaierons pas d'esquisser une démonstration de la convergence des estimateurs comparables à celui de Newey-West. Nous avons fait allusion à la nature des conditions de régularité requises pour la convergence: les matrices d'autocovariance des moments empiriques doivent tendre vers zéro suffisamment vite lorsque  $p$  augmente. La justification théorique de l'estimateur de Newey-West va également bien au-delà du but recherché dans cet ouvrage. Elle repose sur des considérations de ce que l'on appelle "représentation dans le domaine des fréquences" des  $\mathbf{F}_t$  ainsi que sur un nombre de procédures d'estimation non paramétriques associées. Les lecteurs intéressés sont orientés vers Andrews (1991b) pour un traitement assez complet des nombreuses conclusions. Cet article suggère des alternatives à l'estimateur de Newey-West, et montre qu'ils sont préférables dans certaines circonstances. Malgré tout, les performances de l'estimateur de Newey-West ne sont jamais nettement inférieures à celles des estimateurs proposés. Par conséquent, sa simplicité plaide en sa faveur.

Retournons à présent au modèle IV dont les moments empiriques sont donnés par  $\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ . Afin d'être capable d'utiliser (17.64), nous supposons que le véritable aléa  $u_t \equiv y_t - \mathbf{X}_t\boldsymbol{\beta}_0$  satisfait une condition de mixité adéquate. Alors les matrices d'autocovariance empiriques  $\hat{\Gamma}(j)$  pour  $j = 0, \dots, p$ , pour  $p$  donné, se calculent comme suit. Une procédure IV ordinaire permet d'obtenir une estimation préliminaire  $\boldsymbol{\beta}_0$  convergente. Puis les résidus  $\hat{u}_t$  sont combinés aux instruments par produit direct  $\hat{\mathbf{V}} \equiv \hat{\mathbf{u}} * \mathbf{W}$ . Alors  $\hat{\Gamma}(j)$  est  $n^{-1}$  fois la matrice de dimension  $l \times l$  des produits scalaires des colonnes de  $\hat{\mathbf{V}}$  avec ces mêmes colonnes retardées  $j$  fois, en remplaçant les éléments non observés par des zéros. Comme nous l'avons vu précédemment,

$\hat{\Gamma}(0)$  correspond à  $n^{-1}\mathbf{W}^\top \hat{\Omega} \mathbf{W}$ , où  $\hat{\Omega} = \text{diag}(\hat{u}_t^2)$ . Enfin,  $\hat{\Phi}$  est construite à l'aide de (17.64).

Comme précédemment, la matrice  $\hat{\Phi}$  ainsi obtenue peut servir dans deux directions. La première consiste à construire ce que l'on appelle l'estimateur de la matrice de covariance de l'estimateur IV ordinaire **robuste à l'hétéroscédasticité et à l'autocorrélation**, ou estimateur **HAC**. Puisque l'estimateur IV est basé sur les moments empiriques  $\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\beta)$  et sur la matrice de pondération  $(\mathbf{W}^\top \mathbf{W})^{-1}$ , comme on peut le voir dans (17.09), l'estimateur de la matrice de covariance HAC est obtenu en appliquant la formule (17.31) dans ce contexte et en utilisant (17.33) et (17.34). Le résultat est

$$(\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} n \hat{\Phi} (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{P}_W \mathbf{X})^{-1}. \quad (17.65)$$

Dans le cas simple où  $\mathbf{W} = \mathbf{X}$ , cette formule relativement lourde devient

$$(\mathbf{X}^\top \mathbf{X})^{-1} n \hat{\Phi} (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Lorsqu'il n'y a pas d'autocorrélation, ce qui implique que  $n\hat{\Phi} = \mathbf{W}^\top \hat{\Omega} \mathbf{W}$ , nous retrouvons la HCCME (16.15) typique d'un modèle de régression linéaire. Cela serait un bon exercice de voir ce que devient (17.65) en l'absence de corrélation en série lorsque  $\mathbf{W} \neq \mathbf{X}$ .

L'estimateur analogue à l'estimateur H2SLS, (17.44), est encore plus intéressant que l'estimateur de la matrice de covariance HAC. Pour cela, nous n'utilisons plus  $(\mathbf{W}^\top \mathbf{W})^{-1}$  comme matrice de pondération, mais l'inverse de  $\hat{\Phi}$ , calculée selon la procédure précédente à l'aide d'un estimateur IV ordinaire en tant qu'estimateur préliminaire convergent. La fonction critère devient

$$(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{W} \hat{\Phi}^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\beta),$$

et l'estimateur, que l'on appelle quelquefois **estimateur des doubles moindres carrés en deux étapes**, est par conséquent

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{W} \hat{\Phi}^{-1} \mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \hat{\Phi}^{-1} \mathbf{W}^\top \mathbf{y}. \quad (17.66)$$

Cet estimateur est très similaire à (17.44). Dans le cas de ce dernier, la matrice  $\hat{\Phi}$  est remplacée par  $\mathbf{W}^\top \hat{\Omega} \mathbf{W}$ , qui correspond véritablement à l'estimation adéquate de  $\Phi$  en l'absence d'autocorrélation. Il est plus facile d'obtenir une estimation de la matrice de covariance asymptotique de (17.66) plutôt que celle de l'estimateur IV ordinaire. C'est

$$\hat{V}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{W} \hat{\Phi}^{-1} \mathbf{W}^\top \mathbf{X})^{-1}.$$

Il y a eu jusqu'à présent très peu d'expérimentations pratiques sur l'estimateur (17.66). L'une des raisons de ce manque d'intérêt est que les

économètres préfèrent modéliser les dynamiques de façon explicite (voir le Chapitre 19) plutôt que les conserver dans l'aléa et utiliser un estimateur robuste à la spécification. Même si ce dernier fournit des estimations convergentes de certains paramètres, il peut passer sous silence les plus intéressants et provoquer une mauvaise spécification des aléas sans qu'elle soit détectée. Une autre raison est que l'on connaît mal ses propriétés avec des échantillons finis. Les résultats de Cragg (1983) et Tauchen (1986) pour les estimateurs comparables suggèrent qu'elles sont quelquefois pauvres.

Un problème pratique important concerne le choix du paramètre de troncature  $p$ . La théorie est manifestement muette à ce sujet. Ainsi que nous l'avons mentionné, il existe des résultats qui établissent le taux auquel  $p$  doit tendre vers l'infini lorsque  $n$  tend vers l'infini. Mais si l'on dispose d'un échantillon qui contient précisément 136 observations, quelle valeur de  $p$  choisir? Andrews (1991b) s'attaque de front à ce problème et fournit des méthodes de choix pour  $p$  basées sur les données et sur l'estimation d'une valeur optimale d'un paramètre qu'il définit. Il est juste de dire qu'aucune de ses méthodes n'est élémentaire, et nous ne pouvons pas les exposer ici. Le résultat vraisemblablement le plus encourageant de ses recherches est que, au voisinage de la valeur optimale de  $p$ , les variations de  $p$  ont peu d'influence sur les performances de l'estimateur HAC.

Andrews (1991b) fournit également une conclusion appréciable sur les estimateurs des matrices de covariance HAC, (17.64) ainsi que d'autres, à partir d'expériences Monte-Carlo. Le résultat sans doute le plus important est qu'*aucun* des estimateurs HAC qu'il considère n'est fiable pour des tailles d'échantillon inférieures à 250 ou si les aléas obéissent à un processus AR(1) dont le paramètre d'autocorrélation est supérieur à 0.9. Ce résultat décourageant provient du fait que les processus AR(1) avec des paramètres proches de 1 sont comparables à ceux qui possèdent une **racine unitaire**. Ce phénomène est traité dans le Chapitre 20, et nous verrons que les racines unitaires jettent un trouble dans la théorie économétrique traditionnelle.

Si nous nous éloignons des racines unitaires tout en restant proches, les choses sont plus régulières. Nous avons vu au cours du Chapitre 16 qu'il est possible d'employer des HCCME même en présence d'homoscédasticité sans grande perte de précision, à condition d'utiliser l'une des meilleures HCCME. Il apparaît que l'on peut procéder de la même manière pour les HAC. Dans le cas d'un modèle de régression ordinaire avec des aléas indépendants en série et homoscédastiques, la perte de précision due à l'usage de l'estimateur de Newey-West en comparaison de l'estimateur OLS habituel  $\hat{\sigma}^2(\mathbf{X}^\top \mathbf{X})^{-1}$ , par exemple, est faible. Avec quelques uns des autres estimateurs HAC considérés par Andrews, la perte est encore plus faible, ce qui implique que l'estimateur de Newey-West n'est en général pas le meilleur disponible. De façon similaire, si les aléas sont hétéroscédastiques mais indépendants en série, une HCCME est bien meilleure que l'estimateur OLS mais seulement un peu meilleure que l'estimateur HAC.

Si les aléas sont autocorrélés à l'ordre un mais homoscédastiques, aussi bien l'estimateur OLS que la HCCME sont dominés non seulement par l'estimateur HAC, ainsi que l'on pouvait s'y attendre, mais aussi par l'estimateur immédiat calculé en estimant le paramètre d'autocorrélation  $\rho$  et en utilisant l'estimateur de la matrice de covariance d'une procédure de GLS faisables. Ce dernier estimateur est dans ces circonstances préférable aux estimateurs HAC. En réalité, c'est seulement lorsque les aléas sont hétéroscédastiques et autocorrélés que les estimateurs HAC affirment leur supériorité. Même dans de telles circonstances, il est possible, avec certains schémas d'hétéroscédasticité, que l'estimateur GLS, qui ne prend pas en compte une possible hétéroscédasticité, soit préférables aux estimateurs HAC. Mais c'est probablement l'exception plutôt que la règle, puisqu'Andrews trouve d'autres schémas d'hétéroscédasticité qui, combinés à de l'autocorrélation, nécessitent l'usage des estimateurs HAC pour produire une inférence suffisamment précise.

A l'évidence le débat sur les estimateurs HAC n'est pas entièrement clos. Par exemple, dans les exécutions habituelles de l'estimateur de Newey-West pour les modèles IV linéaires,  $\hat{T}(0)$  correspond à  $n^{-1}\mathbf{W}^\top \hat{\Omega} \mathbf{W}$ , où  $\hat{\Omega}$  est l'estimateur relativement pauvre associé à la forme  $HC_0$  de l'HCCME. Il semble raisonnable de penser qu'il serait plus profitable d'employer d'autres formes de  $\Omega$  dans l'estimateur de Newey-West, comme dans les HCCME, et de trouver des moyens similaires d'améliorer les estimateurs  $\hat{T}(j)$  pour  $j \neq 0$ . Cependant, à l'instant où nous écrivons, rien ne permet de croire que ces conjectures sont justifiées. Une approche assez différente, dont nous ne discuterons pas, a été proposée récemment par Andrews et Monahan (1992).

Au cours de la prochaine section, nous abandonnerons les "détails polluants" de l'estimation de la matrice de covariance, en supposant que l'on dispose d'un estimateur adéquat, et reporterons notre attention sur les tests asymptotiques des contraintes de suridentification ainsi que sur d'autres aspects des tests de spécification pour les modèles GMM.

## 17.6 INFÉRENCE DANS LES MODÈLES GMM

Dans cette section, nous proposons une étude des tests d'hypothèses dans un contexte de modèles GMM. Nous débutons par l'examen des tests de contraintes de suridentification, puis développons des procédures qui s'apparentent aux tests classiques étudiés lors du Chapitre 13 pour les modèles estimés par maximum de vraisemblance. Les similitudes avec les procédures déjà étudiées sont frappantes. Il existe une différence importante malgré tout: nous ne pourrons pas faire un usage important des régressions artificielles dans le but d'exécuter les tests dont nous discutons. La raison est simplement que de telles régressions artificielles n'ont pas été développées de façon satisfaisante. Elles existent uniquement dans quelques cas particuliers, et leurs propriétés avec des échantillons de taille finie sont pratiquement inconnues. Cependant, il y a toute raison de croire et d'espérer que dans quelques années, il sera

possible de réaliser des inférences à partir des modèles GMM aux moyens de régressions artificielles qu'il reste à inventer.

En attendant, il existe de nombreuses procédures de tests pour les modèles GMM faciles à exécuter. La plus importante est le test des contraintes de suridentification que l'on impose habituellement. Supposons que l'on ait estimé un vecteur  $\theta$  de  $k$  paramètres en minimisant la fonction critère

$$\iota^\top F(\theta) \hat{\Phi}^{-1} F^\top(\theta) \iota, \quad (17.67)$$

dans laquelle la matrice des moments empiriques  $F(\theta)$  possède  $l > k$  colonnes. Observons que l'on a employé une matrice de pondération  $\hat{\Phi}^{-1}$  qui satisfait la condition nécessaire du Théorème 17.3 pour l'efficacité de l'estimateur GMM. Seules  $k$  conditions sur les moments sont nécessaires pour identifier les  $k$  paramètres, de sorte qu'il y a  $l - k$  contraintes de suridentification implicites dans l'estimation que nous avons exécutée. Comme nous l'avons souligné lors du Chapitre 7, où nous avons rencontré pour la première fois des contraintes de suridentification, il faudrait toujours tester dans la pratique ces contraintes avant de faire un usage quelconque des résultats de l'estimation.

Un moyen de le faire, et qui fut suggéré par Hansen (1982), consiste à employer comme statistique de test la valeur de la fonction critère minimisée. La statistique de test est (17.67) évaluée en  $\theta = \hat{\theta}$  et divisée par la taille de l'échantillon  $n$ :

$$\frac{1}{n} \iota^\top \hat{F} \hat{\Phi}^{-1} \hat{F}^\top \iota, \quad (17.68)$$

où, comme d'habitude,  $\hat{F}$  désigne  $F(\hat{\theta})$ . Le facteur  $n^{-1}$  est nécessaire pour compenser le facteur  $n$  dans  $\hat{\Phi}^{-1}$ , qui apparaît du fait que  $\Phi$  est définie dans (17.29) comme la matrice de covariance de  $n^{-1/2} F_0^\top \iota$ . La définition (17.29) implique par conséquent que si les contraintes de suridentification sont exactes, la distribution asymptotique de  $n^{-1/2} F_0^\top \iota$  est  $N(0, \Phi)$ .

Cependant, pour des raisons qui doivent maintenant nous paraître familières, la distribution asymptotique de  $\hat{F}^\top \iota$  n'est pas la même que la distribution asymptotique de  $F_0^\top \iota$ . Afin d'obtenir une matrice de covariance correcte pour le vecteur en question, nous exécuterons un développement de Taylor en série comme suit:

$$\begin{aligned} n^{-1/2} \hat{F}^\top \iota &\stackrel{a}{=} n^{-1/2} F_0^\top \iota + \frac{1}{n} \sum_{j=1}^k \sum_{t=1}^n \frac{\partial F_t^\top}{\partial \theta_j}(\theta_0) n^{1/2} (\hat{\theta} - \theta_0)_j \\ &\stackrel{a}{=} n^{-1/2} F_0^\top \iota + D(\mu, \theta_0) n^{1/2} (\hat{\theta} - \theta_0). \end{aligned}$$

Posons  $D = D(\mu, \theta_0)$ , et il suit que de (17.22), (17.27), et (17.28),

$$n^{1/2} (\hat{\theta} - \theta_0) \stackrel{a}{=} -(D^\top \Phi^{-1} D)^{-1} D^\top \Phi^{-1} n^{-1/2} F_0^\top \iota.$$

Par conséquent

$$n^{-1/2} \hat{F}^\top \iota \stackrel{a}{=} \left( I - D(D^\top \Phi^{-1} D)^{-1} D^\top \Phi^{-1} \right) n^{-1/2} F_0^\top \iota. \quad (17.69)$$

Soit  $\hat{\Psi}$  une matrice de dimension  $l \times l$  symétrique et définie positive telle que  $\hat{\Psi}^2 = \hat{\Phi}^{-1}$ . Alors la fonction critère minimisée (17.68) devient une norme au carré du vecteur  $n^{-1/2}\hat{\Psi}\hat{F}^\top\boldsymbol{\iota}$ . De (17.69), ce vecteur est asymptotiquement équivalent à

$$\begin{aligned} & \Psi \left( \mathbf{I} - D(D^\top \Psi^2 D)^{-1} D^\top \Psi^2 \right) n^{-1/2} \mathbf{F}_0^\top \boldsymbol{\iota} \\ &= \left( \mathbf{I} - \Psi D(D^\top \Psi^2 D)^{-1} D^\top \Psi \right) \Psi n^{-1/2} \mathbf{F}_0^\top \boldsymbol{\iota} \\ &= \mathbf{M}_{\Psi D} \Psi n^{-1/2} \mathbf{F}_0^\top \boldsymbol{\iota}, \end{aligned}$$

où  $\Psi^2 = \Phi^{-1}$ , et où  $\mathbf{M}_{\Psi D}$  est la matrice de dimension  $l \times l$  qui projette orthogonalement sur le complément orthogonal de l'espace engendré par les  $k$  colonnes de  $\Psi D$ . Par construction, le vecteur  $n^{-1/2}\Psi \mathbf{F}_0^\top \boldsymbol{\iota}$  de dimension  $l$  possède la distribution  $N(\mathbf{0}, \mathbf{I})$ . Il s'ensuit que (17.68) est asymptotiquement distribuée suivant une loi du chi carré dont le nombre de degrés de liberté est égal au rang de  $\mathbf{M}_{\Psi D}$ , soit  $l - k$ , le nombre des contraintes de suridentification.

Le **test des contraintes de suridentification de Hansen** est totalement analogue, dans le contexte plus général actuel, au test pour l'estimation IV dont nous avons discuté dans la Section 7.8, basé sur la fonction critère (7.56). C'est un bon exercice que de faire la dérivation donnée précédemment dans le cas d'un modèle de régression linéaire où les aléas sont homoscédastiques et indépendants en série, afin de voir à quel point le cas général est comparable au cas simple.<sup>2</sup>

Le test des contraintes de suridentification de Hansen est très comparable à ce que l'on connaît en économétrie sous le nom de test de spécification portmanteau. Parce que les modèles estimés par GMM sont soumis à si peu de contraintes, leur "spécification" ne demande pas trop d'efforts. En particulier, si l'on ne réclame pas plus que l'existence des moments employés pour l'identification des paramètres, seuls deux éléments peuvent faire l'objet d'un test. Le premier est l'ensemble de toutes les contraintes de suridentification utilisées, et le second est la constance des paramètres.<sup>3</sup> Parce que le test des contraintes de suridentification de Hansen possède autant de degrés de liberté qu'il y a de contraintes de suridentification, il peut être possible d'obtenir davantage de puissance en diminuant le nombre des degrés de liberté. Cependant, si la statistique de test de Hansen est numériquement assez faible, un tel test ne rejettera jamais l'hypothèse nulle, pour la simple raison que la statistique de Hansen fournit une borne supérieure à toutes les statistiques de test pour lesquelles l'hypothèse nulle correspond au modèle estimé. Cela provient

<sup>2</sup> La statistique de test de Hansen, (17.68), est quelquefois appelée statistique  $J$ . Pour des raisons évidentes (voir le Chapitre 11), nous préférons ne pas lui donner ce nom.

<sup>3</sup> Des tests de constance des paramètres dans des modèles estimés par GMM sont abordés par Hoffman et Pagan (1989) et Ghysels et Hall (1990).

du fait qu'aucune fonction critère du type (17.67) ne peut prendre de valeur négative.

Les tests pour lesquels l'hypothèse nulle n'est pas le modèle estimé ne sont pas soumis à la borne donnée par la statistique de Hansen. Dans le cas contraire, bien évidemment, il deviendrait absolument impossible de rejeter un modèle juste identifié. Un test de constance des paramètres n'est pas soumis non plus à la borne, bien que l'hypothèse nulle semble correspondre à première vue au modèle estimé. La raison fut exposée dans la Section 11.2 en connexion avec les tests de constance des paramètres dans les modèles de régression non linéaire estimés par variables instrumentales. Fondamentalement, afin d'éviter des problèmes d'identification, il est nécessaire de doubler le nombre des instruments employés, en scindant les instruments originaux comme dans (11.09). Les mêmes considérations s'appliquent aux modèles GMM, bien évidemment, et en particulier à ceux qui sont juste identifiés ou qui ont peu de contraintes de suridentification. Mais si l'on emploie deux fois plus d'instruments, le modèle qui correspond à l'hypothèse nulle a été effectivement modifié, et pour cette raison la statistique de Hansen ne donne plus du tout une borne pour les statistiques utilisées lors des tests de constance des paramètres.

Il peut être judicieux de tester d'autres aspects d'un modèle GMM. Dans ces circonstances, ce qui est testé n'est pas tellement la spécification du modèle mais plutôt si des contraintes supplémentaires sur le modèle sont réalistes. Cela suggère l'emploi de tests basés sur le principe de Wald. Supposons donc que nous désirons tester un ensemble de  $r$  contraintes de la forme

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}, \quad \text{où } \mathbf{r} : \Theta \rightarrow \mathbb{R}^r; \quad (17.70)$$

souvenons-nous de (13.02). Le vecteur de paramètres  $\boldsymbol{\theta}$  de dimension  $k$  est défini dans le contexte d'un modèle adéquat, estimé sous sa forme non contrainte par la minimisation de la fonction critère (17.67). Le modèle peut être soit suridentifié, soit juste identifié. Comme d'habitude, nous posons  $\mathbf{R}(\boldsymbol{\theta}) \equiv D_{\boldsymbol{\theta}}\mathbf{r}(\boldsymbol{\theta})$ . Alors, par analogie avec (8.78) et (13.05), nous pouvons construire une statistique de Wald de la façon suivante:

$$W = n\hat{\mathbf{r}}^{\top}(\hat{\mathbf{R}}(\hat{\mathbf{D}}^{\top}\hat{\boldsymbol{\Phi}}^{-1}\hat{\mathbf{D}})^{-1}\hat{\mathbf{R}}^{\top})^{-1}\hat{\mathbf{r}}. \quad (17.71)$$

La justification est exactement la même que celle pour les statistiques Wald et pseudo-Wald vues précédemment: la matrice de covariance asymptotique de  $n^{1/2}\mathbf{r}(\hat{\boldsymbol{\theta}})$  est  $\mathbf{R}(\mathbf{D}^{\top}\boldsymbol{\Phi}^{-1}\mathbf{D})^{-1}\mathbf{R}^{\top}$ . Les difficultés relatives à ce test sont également les mêmes que celles associées aux autres tests de Wald, à savoir que la statistique n'est pas invariante à une reparamétrisation des contraintes. Par conséquent, la statistique (17.71) est généralement peu recommandée et devrait être employée avec précaution si l'on est absolument contraint d'y avoir recours.

Il est aussi envisageable de baser des tests de modèles estimés par GMM sur les principes LM et LR. Pour un test LM, nous exécuterons seulement une estimation contrainte, en minimisant (17.67) sous les contraintes (17.70), pour obtenir les estimations contraintes  $\hat{\theta}$ . Le test LM classique se base sur le gradient de la fonction de logvraisemblance, évalué avec les estimations contraintes. La fonction de logvraisemblance est une fonction critère, il est donc naturel de baser un test LM dans ce contexte sur le gradient de la fonction critère (17.67). Il est aisé de voir que ce gradient est asymptotiquement proportionnel au vecteur aléatoire de dimension  $k$

$$n^{-1/2} \mathbf{D}^\top \Phi^{-1} \mathbf{F}^\top \boldsymbol{\iota}.$$

Ce vecteur est asymptotiquement normal lorsqu'il est évalué en  $\theta_0$ , son espérance est nulle et sa matrice de covariance est

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{D}^\top \Phi^{-1} \mathbf{D} \right),$$

ce qui suggère qu'une statistique de test appropriée serait

$$LM = \frac{1}{n} \boldsymbol{\iota}^\top \tilde{\mathbf{F}} \tilde{\Phi}^{-1} \tilde{\mathbf{D}} (\tilde{\mathbf{D}}^\top \tilde{\Phi}^{-1} \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^\top \tilde{\Phi}^{-1} \tilde{\mathbf{F}}^\top \boldsymbol{\iota}, \quad (17.72)$$

où  $\tilde{\mathbf{D}}$  est défini par (17.32) avec  $\tilde{\theta}$  à la place de  $\hat{\theta}$ ,  $\tilde{\mathbf{F}} \equiv \mathbf{F}(\tilde{\theta})$ , et où  $\tilde{\Phi}$  est un estimateur adéquat de  $\Phi$ ; à la fin de la section précédente, nous promettons de ne pas détailler le calcul de  $\tilde{\Phi}$ .

Il est assez facile de montrer que, sous l'hypothèse nulle, la statistique  $LM$  donnée par (17.72) est distribuée suivant une loi du chi carré à  $r$  degrés de liberté. Il est plus intéressant de montrer que, lorsque le modèle non contraint est juste identifié, (17.72) est numériquement identique à la statistique (17.68) asymptotiquement distribuée selon une chi carré pour les contraintes de suridentification, à condition que le même estimateur de  $\Phi$  soit employé dans les deux statistiques. En réalité, cela provient du fait que la matrice  $\mathbf{D}$  est carrée et non singulière pour des modèles juste identifiés. Puisque  $\mathbf{D}^{-1}$  existe, on peut simplifier l'écriture de (17.72) et obtenir

$$\frac{1}{n} \boldsymbol{\iota}^\top \tilde{\mathbf{F}} \tilde{\Phi}^{-1} \tilde{\mathbf{F}}^\top \boldsymbol{\iota}. \quad (17.73)$$

Cette statistique est identique à (17.68), puisque le vecteur  $\hat{\theta}$  employé est ici une estimation *contrainte*, issue de l'estimation soumise aux contraintes de suridentification.

Notons que (17.72) ne peut pas être numériquement plus grande que (17.73) et sera en général plus faible. Ceci est un exemple supplémentaire de la borne dont nous avons parlé. Nous pouvons voir cela aisément en écrivant (17.72) sous la forme

$$\frac{1}{n} \boldsymbol{\iota}^\top \tilde{\mathbf{F}} \tilde{\Psi} \tilde{\Psi} \tilde{\mathbf{D}} (\tilde{\mathbf{D}}^\top \tilde{\Psi} \tilde{\Psi} \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^\top \tilde{\Psi} \tilde{\Psi} \tilde{\mathbf{F}}^\top \boldsymbol{\iota}$$



et (17.73) sous la forme

$$\frac{1}{n} \boldsymbol{\iota}^\top \tilde{\mathbf{F}} \tilde{\boldsymbol{\Psi}} \tilde{\boldsymbol{\Psi}}^\top \tilde{\mathbf{F}}^\top \boldsymbol{\iota}.$$

Ainsi (17.73) est assimilable à la norme au carré du vecteur  $n^{-1/2} \tilde{\boldsymbol{\Psi}} \tilde{\mathbf{F}}^\top \boldsymbol{\iota}$ , et (17.72) est assimilable à la norme au carré de ce même vecteur après qu'il ait été projeté sur le sous-espace engendré par les colonnes de  $\tilde{\boldsymbol{\Psi}} \tilde{\mathbf{D}}$ .

La statistique  $LR$  pour les modèles GMM a la même simplicité que pour les modèles estimés par maximum de vraisemblance. Elle correspond simplement à la différence entre les valeurs de la fonction critère (17.68) évaluée avec les estimations contraintes et non contraintes:

$$LR = \frac{1}{n} (\boldsymbol{\iota}^\top \tilde{\mathbf{F}} \tilde{\boldsymbol{\Phi}}^{-1} \tilde{\mathbf{F}}^\top \boldsymbol{\iota} - \boldsymbol{\iota}^\top \hat{\mathbf{F}} \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{F}}^\top \boldsymbol{\iota}). \quad (17.74)$$

Ce résultat semble a priori trop beau pour être valable. Après tout, même dans un contexte classique, un facteur de 2 est nécessaire pour la forme  $LR$  du test. La clé de ce résultat est l'hypothèse cruciale que la matrice de pondération employée dans la fonction critère satisfait la condition d'efficacité du Théorème 17.3. Sans cette hypothèse, comme nous le verrons brièvement à la fin de cette section, les choses peuvent se compliquer. Remarquons que  $\tilde{\boldsymbol{\Phi}}$  et  $\hat{\boldsymbol{\Phi}}$  seront souvent identiques à (17.74), parce que s'il est difficile d'estimer  $\boldsymbol{\Phi}$ , il est judicieux de ne l'estimer qu'une seule fois.

Nous ne démontrerons pas la validité de (17.74). Cependant, au moins un cas particulier montre que cette statistique  $LR$  est plausible. Lorsqu'un modèle est juste identifié, la fonction critère a valeur nulle: les  $k$  conditions portant sur les moments empiriques peuvent être satisfaites exactement avec  $k$  paramètres. La différence des fonctions critère est simplement la fonction contrainte, et cela correspond, ainsi que nous l'avons vu, à la statistique de Hansen et à la statistique  $LM$  dans ces circonstances.

Enfin, considérons les tests  $C(\alpha)$ . Soit  $\boldsymbol{\theta}$  un vecteur de paramètres satisfaisant les contraintes  $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ . Alors la statistique de test peut être élaborée comme s'il s'agissait de la différence de deux statistiques  $LM$ , l'une correspondant au modèle contraint et l'autre au modèle non contraint, évaluées toutes deux en  $\boldsymbol{\theta}$ . Supposons, pour simplifier, que le vecteur de paramètres  $\boldsymbol{\theta}$  puisse être partitionné en  $[\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2]$  et que l'on puisse écrire les contraintes sous la forme  $\boldsymbol{\theta}_2 = \mathbf{0}$ . Le premier terme de la statistique  $C(\alpha)$  est de la forme de (17.72) mais il est évalué avec  $\boldsymbol{\theta}$  plutôt qu'avec le véritable estimateur contraint  $\tilde{\boldsymbol{\theta}}$ . Le second terme devrait avoir la forme d'une statistique  $LM$  appropriée au modèle contraint, pour lequel seul  $\boldsymbol{\theta}_1$  peut varier. Cela correspond au remplacement de la matrice  $\tilde{\mathbf{D}}$  dans (17.72) par  $\hat{\mathbf{D}}_1$ , où la partition de  $\mathbf{D}$  en  $[\mathbf{D}_1 : \mathbf{D}_2]$  correspond à la partition de  $\boldsymbol{\theta}$ . Par conséquent, la statistique  $C(\alpha)$  est

$$\begin{aligned} C(\alpha) = & \frac{1}{n} \boldsymbol{\iota}^\top \hat{\mathbf{F}} \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{D}} (\hat{\mathbf{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{D}})^{-1} \hat{\mathbf{D}}^\top \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{F}}^\top \boldsymbol{\iota} \\ & - \frac{1}{n} \boldsymbol{\iota}^\top \hat{\mathbf{F}} \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{D}}_1 (\hat{\mathbf{D}}_1^\top \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{D}}_1)^{-1} \hat{\mathbf{D}}_1^\top \hat{\boldsymbol{\Phi}}^{-1} \hat{\mathbf{F}}^\top \boldsymbol{\iota}. \end{aligned} \quad (17.75)$$

Ici, comme auparavant,  $\hat{\Phi}$  est une estimation adéquate de  $\Phi$ . Afin de montrer que (17.75) est asymptotiquement équivalente à la véritable statistique  $LM$ , il suffit de modifier les détails de démonstration de l'équivalence asymptotique correspondante dans la Section 13.7.

Dans le cas général où les contraintes s'expriment sous la forme  $\mathbf{r}(\theta) = \mathbf{0}$ , une autre forme du test  $C(\alpha)$  peut se révéler plus pratique, puisque la construction d'une matrice correspondant à  $\mathbf{D}_1$  peut être compliquée. Cette forme est

$$\iota^\top \mathbf{F} \hat{\Phi}^{-1} \mathbf{D} (\mathbf{D}^\top \hat{\Phi}^{-1} \mathbf{D})^{-1} \hat{\mathbf{R}}^\top \left( \hat{\mathbf{R}} (\mathbf{D}^\top \hat{\Phi}^{-1} \mathbf{D})^{-1} \hat{\mathbf{R}}^\top \right)^{-1} \hat{\mathbf{R}} (\mathbf{D}^\top \hat{\Phi}^{-1} \mathbf{D})^{-1} \mathbf{D}^\top \hat{\Phi}^{-1} \mathbf{F}^\top \iota.$$

Pour que cette statistique soit pertinente, la difficulté de calcul des estimations contraintes  $\theta$  doit l'emporter sur la difficulté de la formule précédente. La formule elle-même peut être établie, au prix de quelques manipulations algébriques ennuyeuses, en adoptant les méthodes de la Section 8.9. Nous laissons tous ces détails au lecteur intéressé.

Le traitement que nous avons donné des tests LM, LR et Wald suit assez fidèlement celui de Newey et West (1987b). Cet article peut être intéressant à consulter pour davantage de détails sur les conditions de régularité suffisant pour que les résultats soient valables. L'article de Newey (1985b) est également consacré aux tests de modèles estimés par GMM. Les tests d'hypothèses non emboîtées pour les modèles estimés par GMM sont abordés par Smith (1992). Cependant, ces articles ne discutent pas des tests  $C(\alpha)$ .

Une question intéressante est de savoir si les tests de moments conditionnels discutés dans le chapitre précédent dans un contexte de modèles estimés par maximum de vraisemblance ont un équivalent quelconque pour les modèles estimés par GMM. Pour simplifier, supposons qu'il n'y ait qu'un seul moment conditionnel dont l'espérance est nulle si le modèle est correctement spécifié. Si le moment empirique correspondant est employé comme contrainte, alors il peut être testé de la même manière que n'importe quelle autre contrainte, par l'une des procédures décrites précédemment.

Une autre possibilité consiste en un moment resté inemployé pour l'identification ou la suridentification des paramètres du modèle, tel qu'un moment généré par un instrument qui, bien qu'appartenant à l'ensemble d'informations adéquat, n'est pas employé en tant qu'instrument dans la procédure d'estimation. Il est aisé en principe de voir comment construire un test de moment conditionnel dans ce cas. Le modèle doit être estimé à nouveau en utilisant le moment conditionnel qui doit être testé comme contrainte de suridentification. Dans la pratique, cela est plus facile à dire qu'à faire, parce que la matrice  $\Phi$  doit être augmentée d'une ligne et d'une colonne pour ce nouveau moment. La différence entre les deux fonctions critères minimisées, avec et sans le moment supplémentaire, génère la statistique de test  $LR$ .

La raison sous-jacente pour laquelle les tests de moments conditionnels sont, du moins potentiellement, plus délicats à exécuter dans un contexte

GMM que dans un contexte de maximum de vraisemblance est l'absence de méthode basée sur une régression artificielle. Cela est relié à la difficulté d'obtenir des estimations de la matrice  $\Phi$  si nous voulons imposer aussi peu de structure que possible à nos modèles. Pour ces cas où nous imposons suffisamment de contraintes pour constater avec joie que l'estimation de  $\Phi$  est aisée, les tests de moment conditionnel ne sont pas plus difficiles à mettre en oeuvre que dans un contexte de spécification complète du maximum de vraisemblance.

Nous avons limité notre attention dans cette section aux modèles estimés par la minimisation de fonctions critère avec des matrices de pondération satisfaisant la condition d'efficacité du Théorème 17.3. La principale justification de ce choix est que, même si une matrice de pondération non efficace peut quelquefois être adéquate pour des besoins d'estimation, les procédures de test ne peuvent pas être mises en oeuvre sans une estimation de la matrice de covariance  $\Phi$  des moments empiriques, quelle que soit la matrice de pondération utilisée. Il est par conséquent peu pertinent de baser des inférences sur des estimations non efficaces lorsque le travail difficile d'estimation efficace de  $\Phi$  a été réalisé. Une autre raison est que, tout simplement, la théorie des tests basés sur des estimations non efficaces des paramètres est substantiellement plus difficile que la théorie présentée ici.

## 17.7 CONCLUSION

La théorie asymptotique sous-jacente à la méthode des moments généralisée est en réalité assez générale. Elle possède l'attrait des théories qui manipulent des éléments apparemment très variés et qui fournissent un traitement unifié. Nous avons vu au cours de ce chapitre comment chaque estimateur considéré jusqu'à présent peut être compris comme un estimateur GMM, et dans bien des cas, nous avons donné une extension des procédures d'estimation en adoptant un point de vue GMM, les rendant robustes à une plus grande variété de spécifications.

Par souci de simplicité, tous les exemples d'estimateurs GMM présentés dans ce chapitre ont été considérés dans un contexte de modèles linéaires. Il est important de souligner que cela ne constitue en rien une limitation de la méthode. L'extension de nos simples exemples à des cas de régressions non linéaires est entièrement immédiate, du moins théoriquement. Dans la pratique, évidemment, tout, excepté l'estimation GMM la plus simple, doit être mis en oeuvre dans la minimisation numérique de la fonction critère, avec toutes les difficultés habituelles que cela implique. Malgré ces difficultés, l'application majeure des GMM est l'objet de modèles non linéaires.

Jusqu'ici, il est impossible de prévoir dans quelle mesure les GMM modifieront la pratique de l'économétrie. Les tests sont, comme nous l'avons vu, souvent plus difficiles dans une modélisation GMM que dans n'importe

quelle autre catégorie de modèle étudié. Un autre point sur lequel nous restons relativement muets concerne les propriétés des estimateurs GMM et des statistiques de test lorsque l'échantillon a une taille comparable à celle des échantillons concrets. Il est incontestable que des recherches ultérieures clarifieront un grand nombre de ces questions. Nous trouverons une application de la GMM dans le chapitre suivant qui traite des modèles d'équations simultanées.

## TERMES ET CONCEPTS

application définissant des paramètres	instruments optimaux
autocovariances (des moments empiriques)	$M$ -estimateur de Type 2
borne GMM	$M$ -estimateurs
condition sur le moment	matrice d'autocovariance empirique
doubles moindres carrés en deux étapes	matrice de pondération
équation définissant de l'estimateur	méthode des moments généralisée (GMM)
estimateur de la matrice de covariance robuste à l'hétéroscédasticité et à l'autocorrélation (HAC)	méthode des moments (ordinaire)
estimateur GMM	modèle de localisation
estimateur H2SLS (doubles moindres carrés en deux étapes)	moments empiriques
estimateur HOLS	paramètre de troncature des retards
fonction critère	poids optimaux
identifiabilité asymptotique forte	tests $C(\alpha)$ pour modèles GMM
	tests de Wald pour les modèles GMM
	tests des contraintes de suridentification de Hansen
	tests LM pour les modèles GMM
	tests LR pour les modèles GMM

# Chapitre 18

## Modèles d'Equations Simultanées

### 18.1 INTRODUCTION

Pendant de nombreuses années, le **modèles d'équations simultanées linéaire** a été le centre d'intérêt de la théorie économétrique. Nous avons abordé un cas particulier de ce modèle, un modèle d'offre-demande à deux équations, dans la Section 7.3. L'objet de cette discussion était simplement de montrer que la simultanéité implique une corrélation entre les régresseurs et les termes d'erreur de chaque équation de système, rendant les OLS non convergents et justifiant l'usage des variables instrumentales. La non convergence des estimateurs par moindres carrés des équations individuelles dans les modèles d'équations simultanées n'est pourtant pas le seul résultat économétrique pour ce genre de modèle. Dans ce chapitre, nous discutons donc des modèles d'équations simultanées en détail.

La grande majorité du travail récent sur les modèles d'équations simultanées s'est développé sous la bienveillance de la Commission Cowles; Koopmans (1950) et Hood et Koopmans (1953) sont des références connues. Ce travail a fortement influencé la direction suivie par la théorie économétrique depuis de nombreuses années. Pour une histoire sur le développement récent de l'économétrie, consulter Morgan (1990). Parce que la littérature consacrée aux modèles d'équations simultanées est vaste, nous ne traiterons qu'une petite partie de celle-ci. Il existe un grand nombre d'études sur ce champ théorique, et de nombreux ouvrages qui se situent à des niveaux différents. Deux articles de synthèse intéressants sont ceux de Hausman (1983), qui traite de la littérature traditionnelle, et Phillips (1983), qui traite du champ plus spécifique de la théorie en petit échantillon dans les modèles d'équations simultanées, un sujet que nous n'aborderons pas du tout.

La caractéristique essentielle des modèles d'équations simultanées est que deux ou plusieurs **variables endogènes** sont déterminées simultanément par le modèle, comme des fonctions de **variables exogènes**, de **variables prédéterminées**, et d'aléas. A ce stade, nous en avons dit très peu sur ce que nous entendons par variables exogènes et prédéterminées. Puisque le rôle de telles variables est essentiel dans les modèles d'équations simultanées, il est temps de corriger le défaut. Dans la Section 18.2, nous discutons par conséquent en détail du concept important de l'**exogénéité**.

La majeure partie du chapitre sera consacrée au modèle d'équations simultanées. Supposons qu'il y ait  $g$  variables endogènes, et par conséquent  $g$  équations, et  $k$  variables exogènes ou prédéterminées. Alors le modèle peut être écrit sous forme matricielle comme

$$\mathbf{Y}\mathbf{\Gamma} = \mathbf{X}\mathbf{B} + \mathbf{U}. \quad (18.01)$$

Ici,  $\mathbf{Y}$  désigne une matrice de dimension  $n \times g$  de variables endogènes,  $\mathbf{X}$  désigne une matrice de dimension  $n \times k$  de variables exogènes ou prédéterminées,  $\mathbf{\Gamma}$  désigne une matrice de dimension  $g \times g$  de coefficients,  $\mathbf{B}$  désigne une matrice de dimension  $k \times g$  de coefficients, et  $\mathbf{U}$  désigne une matrice de dimension  $n \times g$  de termes d'erreur.

Il est immédiatement clair que le modèle (18.01) comprend beaucoup trop de paramètres à estimer. Une observation type pour l'équation  $l$  peut s'écrire sous la forme

$$\sum_{i=1}^g \Gamma_{il} Y_{ti} = \sum_{j=1}^k B_{jl} X_{tj} + u_{tl}.$$

La multiplication de tous les paramètres  $\Gamma_{il}$  et  $B_{jl}$  par n'importe quelle constante non nulle aurait pour effet de multiplier  $u_{tl}$  par cette constante pour tout  $t$ , mais ne modifierait pas la structure des aléas dans les observations. Il est donc nécessaire d'imposer une sorte de normalisation pour chaque équation du modèle. Une normalisation évidente consiste à poser  $\Gamma_{ii} = 1$  pour tout  $i$ ; chaque variable endogène, de  $y_1$  à  $y_g$ , serait alors associée à un coefficient unitaire dans une et une seule équation. Cependant, comme nous l'avons vu dans la Section 7.3, de nombreuses autres normalisations pourraient être envisagées. Nous pourrions, par exemple, poser  $\Gamma_{1l} = 1$  pour tout  $l$ ; le coefficient associé à la première variable endogène serait ainsi égal à l'unité dans chaque équation.

Le modèle (18.01) n'a pas de sens si la matrice  $\mathbf{\Gamma}$  n'est pas inversible, car sinon il serait impossible de déterminer  $\mathbf{Y}$  de manière unique en tant que fonction de  $\mathbf{X}$  et  $\mathbf{U}$ . Nous pouvons donc postmultiplier des deux membres de (18.01) par  $\mathbf{\Gamma}^{-1}$  pour obtenir

$$\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{\Gamma}^{-1} + \mathbf{U}\mathbf{\Gamma}^{-1} \quad (18.02)$$

$$= \mathbf{X}\mathbf{\Pi} + \mathbf{V}. \quad (18.03)$$

L'expression (18.02) est la **forme réduite contrainte**, ou **FRC**, et l'expression (18.03) est la **forme réduite libre**, ou **FRL**. Les contraintes sont  $\mathbf{\Pi} = \mathbf{B}\mathbf{\Gamma}^{-1}$ . Notons que, même dans le cas improbable où les colonnes de  $\mathbf{U}$  étaient indépendantes, celles de  $\mathbf{V}$  ne le seraient pas. Ainsi les diverses équations de la forme réduite possèdent presque sûrement des aléas corrélés.

L'imposition des contraintes de normalisation est nécessaire mais non suffisante pour obtenir des estimations de  $\mathbf{\Gamma}$  et  $\mathbf{B}$ . Le problème est que, à

moins de lui imposer des contraintes, le modèle (18.01) a beaucoup trop de paramètres inconnus. La matrice  $\mathbf{\Gamma}$  possède  $g^2 - g$  coefficients, du fait des  $g$  contraintes de normalisation, alors que la matrice  $\mathbf{B}$  en possède  $gk$ . Il y a donc  $g^2 + gk - g$  coefficients structurels au total. Mais la matrice  $\mathbf{\Pi}$  sous la forme réduite libre ne possède que  $gk$  coefficients. Il est à l'évidence impossible de déterminer les  $g^2 + gk - g$  coefficients structurels à partir des  $gk$  coefficients de la FRL. Il faudra imposer *au moins*  $g^2 - g$  contraintes sur  $\mathbf{\Gamma}$  et/ou  $\mathbf{B}$  afin d'être en mesure d'identifier le modèle. Il existe une vaste littérature consacrée à l'identification dans les modèles d'équations simultanées, qui aborde le problème des conditions sous lesquelles certains ou tous les paramètres de tel modèle peuvent être identifiés. Nous livrerons les principaux résultats de cette littérature dans la Section 18.3.

La grande partie restante du chapitre traite des méthodes d'estimation diverses et variées pour les modèles d'équations simultanées. La Section 18.4 aborde l'estimation par maximum de vraisemblance du modèle dans son ensemble sous l'hypothèse de normalité, une technique connue sous le nom de **maximum de vraisemblance en information complète**, ou **FIML**. La section qui suit traite de l'estimation par maximum de vraisemblance de chaque équation séparément, technique que l'on nomme **maximum de vraisemblance en information limitée**, ou **LIML**. Puis dans la Section 18.6, nous discuterons des **triples moindres carrés**, ou **3SLS**, que l'on dérive comme une application de la méthode des moments généralisée. Enfin, les modèles d'équations simultanées seront abordés dans la Section 18.7.

## 18.2 EXOGÉNÉITÉ ET CAUSALITÉ

Dans le cas d'une équation de régression unique, nous estimons la distribution, ou du moins l'espérance et la variance, d'une variable endogène *conditionnellement* aux valeurs de certaines variables explicatives. Dans le cas d'un modèle d'équations simultanées, nous estimons la distribution jointe de deux ou plusieurs variables endogènes *conditionnellement* aux valeurs de certaines variables explicatives. Mais nous n'avons encore rien dit sur les conditions sous lesquelles nous pouvons considérer une variable comme explicative. Pour que l'inférence conditionnelle soit valable, les variables explicatives doivent être soit **prédéterminées** soit **exogènes** dans un sens ou un autre que nous allons définir.

Dans un contexte de série temporelle, nous avons vu que les variables aléatoires qui sont prédéterminées peuvent être employées sans risque en tant que variables explicatives dans une estimation par moindres carrés, du moins asymptotiquement. En réalité, les variables endogènes retardées sont abondamment utilisées en tant que variables explicatives et en tant qu'instruments. Cependant, il y a de nombreux cas, et parmi eux le cas des modèles estimés à l'aide de données en coupe transversale, où nous voulons utiliser en tant que variables explicatives des variables qui ne sont pas des variables

prédéterminées. De plus, le concept de prédétermination se révèle être plus délicat que ce que l'on imagine, puisque la prédétermination n'est pas invariante à la paramétrisation du modèle. Ainsi il est calir que nous avons besoin d'un concept plus général que celui de la prédétermination.

Il est pratique de débiter par des définitions formelles du concept de prédétermination et du concept étroitement relié de l'**exogénéité faible**. Ce faisant, nous suivons l'exposé classique de ces thèmes, tel qu'il apparaît chez Engle, Hendry, et Richard (1983). Les lecteurs devraient être prévenus que cet article, bien qu'étant une référence classique, n'est pas du tout évident à lire. Notre discussion sera grandement simplifiée par rapport à la leur, et se fondera sur un contexte plus général, puisque ces auteurs se concentrent sur les modèles paramétriques pleinement spécifiés et estimables par maximum de vraisemblance. Nous nous référerons, malgré tout, à un de leurs exemples pour une illustration concrète d'un nombre de points.

Soit  $\mathbf{Y}_t$  le vecteur de dimension  $1 \times g$  l'observation  $t$  d'un ensemble de variables que nous voulons modéliser dans un processus simultané, et soit  $\mathbf{X}_t$  le vecteur de dimension  $1 \times k$  l'observation  $t$  d'un ensemble de variables explicatives, dont toutes ou certaines peuvent être des  $\mathbf{Y}_t$  retardés. Nous pouvons écrire un modèle d'équations simultanées, en général non linéaire, sous la forme

$$\mathbf{h}_t(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta}) = \mathbf{U}_t, \quad (18.04)$$

où  $\mathbf{h}_t$  est un vecteur de dimension  $1 \times g$  de fonctions, comparable à la fonction de régression d'un modèle univarié, où  $\boldsymbol{\theta}$  est un vecteur de paramètres de dimension  $p$ , et où  $\mathbf{U}_t$  est un vecteur de dimension  $1 \times g$  d'aléas. Le modèle linéaire (18.01) peut être considéré comme un cas particulier de (18.04) si nous le mettons sous la forme

$$\mathbf{Y}_t \boldsymbol{\Gamma} = \mathbf{X}_t \mathbf{B} + \mathbf{U}_t$$

et si nous faisons en sorte que  $\boldsymbol{\theta}$  soit composé de tous les éléments de  $\boldsymbol{\Gamma}$  et  $\mathbf{B}$  qu'il faut estimer. Ici  $\mathbf{X}_t$  et  $\mathbf{Y}_t$  sont les  $t^{\text{ième}}$  lignes des matrices  $\mathbf{X}$  et  $\mathbf{Y}$ . On pourrait baser un ensemble de conditions portant sur les moments (conditionnels) sur (18.04), en écrivant

$$E(\mathbf{h}_t(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta})) = \mathbf{0},$$

où l'espérance pourrait s'interpréter comme étant conditionnelle à un ensemble d'information approprié.

*Définition 18.1.*

Les variables explicatives  $\mathbf{X}_t$  sont **prédéterminées** dans l'équation  $i$  du modèle (18.04), pour  $i = 1, \dots, g$ , si, pour tout  $t = 1, \dots, n$ ,

$$\mathbf{X}_t \perp\!\!\!\perp u_{i,t+s} \quad \text{pour tout } s \geq 0.$$



Le symbole  $\perp\!\!\!\perp$  est ici employé pour exprimer l'indépendance statistique. La définition est valable quel que soit le contexte, et en particulier le contexte des séries temporelles pour lequel il existe un ordre naturel. Le prochain concept ne nécessite pas un tel ordonnancement.

*Définition 18.2.*

Les variables explicatives  $\mathbf{X}_t$  sont **strictement exogènes** dans l'équation  $i$  du modèle (18.04) si, pour tout  $t = 1, \dots, n$ ,

$$\mathbf{X}_t \perp\!\!\!\perp \mathbf{U}_s \quad \text{pour tout } s = 1, \dots, n.$$

Si (18.04) représente une forme structurelle, alors autant la prédétermination que l'exogénéité stricte nous autorise à traiter cette forme comme une caractérisation du processus générant  $\mathbf{Y}_t$  conditionnellement à  $\mathbf{X}_t$ . Ainsi nous pouvons, par exemple, écrire une fonction de log-vraisemblance basée sur (18.04), que l'on peut maximiser pour obtenir des estimations convergentes des paramètres  $\boldsymbol{\theta}$ ; voir la Section 18.4. Si l'on pense que (18.04) doit fournir des conditions portant sur les moments conditionnels, alors autant la prédétermination que l'exogénéité stricte nous autorise à employer les colonnes de  $\mathbf{X}$  comme instruments dans l'estimation de  $\boldsymbol{\theta}$  par une sorte quelconque de procédure IV, telle que les 2SLS, 3SLS ou la GMM. En réclamant cette propriété, nous supposons qu'il y a suffisamment d'instruments dans  $\mathbf{X}$  pour *identifier* tous les paramètres de  $\boldsymbol{\theta}$ .

Hélas, le concept de l'exogénéité stricte est beaucoup trop contraignant, du moins pour les applications sur séries temporelles. Dans ce contexte, un très petit nombre de variables sont strictement exogènes, bien que beaucoup soient prédéterminées. Cependant, comme nous allons le montrer, une variable peut être prédéterminée ou non dans un même modèle selon la manière de le paramétrer. En plus de cela, la prédétermination n'est pas toujours nécessaire pour une estimation convergente. Ce concept est par conséquent très peu satisfaisant.

Considérons le modèle simultané suivant, tiré de Engle, Hendry, et Richard (1983):

$$y_t = \beta x_t + \varepsilon_{1t} \tag{18.05}$$

$$x_t = \delta_1 x_{t-1} + \delta_2 y_{t-1} + \varepsilon_{2t}, \tag{18.06}$$

où les aléas sont normalement, identiquement, et indépendamment distribués pour tout  $t$ , avec une matrice de covariance donnée par

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

Si  $\sigma_{12} \neq 0$ ,  $x_t$  est corrélé à  $\varepsilon_{1t}$  et l'estimation de (18.05) par OLS ne sera pas convergente parce que  $x_t$  n'est pas prédéterminé dans (18.05).

Considérons à présent l'espérance de  $y_t$  conditionnellement à  $x_t$  et à tous les  $y_t$  et  $x_t$  retardés. Nous avons

$$E(y_t | x_t, y_{t-1}, x_{t-1} \dots) = \beta x_t + E(\varepsilon_{1t} | x_t, y_{t-1}, x_{t-1} \dots). \quad (18.07)$$

Remarquons que  $\varepsilon_{2t}$  est défini par (18.06) comme une combinaison linéaire des variables conditionnantes. Ainsi l'espérance conditionnelle de  $\varepsilon_{1t}$  dans (18.07) est

$$E(\varepsilon_{1t} | \varepsilon_{2t}) = \frac{\sigma_{12}}{\sigma_{22}} \varepsilon_{2t} = \frac{\sigma_{12}}{\sigma_{22}} (x_t - \delta_1 x_{t-1} - \delta_2 y_{t-1}).$$

Nous pouvons par conséquent écrire

$$y_t = \beta x_t + c_1 x_{t-1} + c_2 y_{t-1} + v_t, \quad (18.08)$$

avec

$$b = \beta + \frac{\sigma_{12}}{\sigma_{22}}, \quad c_1 = -\delta_1 \frac{\sigma_{12}}{\sigma_{22}}, \quad c_2 = -\delta_2 \frac{\sigma_{12}}{\sigma_{22}}, \quad (18.09)$$

où  $v_t$  est indépendant de  $x_t$ . Ainsi  $x_t$  est prédéterminé dans (18.08), quelle que soit la valeur de  $\sigma_{12}$ , bien qu'il ne soit pas prédéterminé dans (18.05) lorsque  $\sigma_{12} \neq 0$ .

Nous retournerons à ce modèle plus tard. Pendant ce temps, progressons vers un concept plus approprié que la prédétermination dans le contexte du modèle simultané. Parce que nous voulons savoir si les variables explicatives  $\mathbf{X}_t$  sont déterminées simultanément aux  $\mathbf{Y}_t$  nous aurons besoin de travailler avec des DGP qui génèrent à la fois  $\mathbf{Y}_t$  et  $\mathbf{X}_t$ . Comme d'habitude, nous pouvons représenter un DGP par une densité de probabilité, ou mieux par son logarithme, que l'on peut exprimer comme la somme de contributions de chaque observation; voir la Section 8.2. La contribution de l'observation  $t$  est de la forme

$$\ell_t(\mathbf{Y}_t, \mathbf{X}_t | \Omega_t). \quad (18.10)$$

Cette expression est le logarithme de la densité jointe de  $\mathbf{Y}_t$  et  $\mathbf{X}_t$  conditionnellement à l'ensemble d'information  $\Omega_t$ . Ce dernier est composé de toutes les observations sur  $\mathbf{Y}_t$  et  $\mathbf{X}_t$ , de la première à la  $(t-1)^{\text{th}}$ .

L'expression (18.10) peut être décomposée en deux contributions, l'une correspondant au logarithme de la densité de  $\mathbf{Y}_t$  conditionnellement à  $\mathbf{X}_t$  et  $\Omega_t$ , et la seconde correspondant au logarithme de la densité de  $\mathbf{X}_t$  conditionnellement à  $\Omega_t$ :

$$\ell_t(\mathbf{Y}_t, \mathbf{X}_t | \Omega_t) = \ell_t^Y(\mathbf{Y}_t | \mathbf{X}_t, \Omega_t) + \ell_t^X(\mathbf{X}_t | \Omega_t), \quad (18.11)$$

avec une notation évidente. A ce stade, nous souhaitons pouvoir faire abstraction de la seconde partie des contributions dans (18.11), puisqu'elle ne concerne que les variables explicatives.

Sous quelles conditions pouvons-nous faire abstraction de la seconde contribution? Pour répondre à cette question, considérons tout d'abord un modèle,  $\mathbb{M}$ , composé de DGP représentés par des ensembles de contributions de la forme (18.11). Puis, définissons une application définissante des paramètres:  $\mathbb{M} \rightarrow \Theta \in \mathbb{R}^p$  qui associe un vecteur de paramètres à  $p$  composantes  $\theta(\mu) \in \Theta$  à chaque  $\mu \in \mathbb{M}$ . Le vecteur de paramètres  $\theta$  contient les **paramètres d'intérêt**, c'est-à-dire ceux que nous voulons estimer. Comme nous allons le voir, il peut y avoir d'autres paramètres, appelés, **paramètres perturbateurs**, que nous ne souhaitons pas estimer.

*Définition 18.3.*

Les variables explicatives  $\mathbf{X}_t$  sont **faiblement exogènes** pour le modèle paramétrique  $(\mathbb{M}, \theta)$  si

- (i) il existe un sous-modèle  $\mathbb{M}^X$  qui contient les DGP pour les variables explicatives  $\mathbf{X}_t$  seulement;
- (ii) il existe un sous-modèle conditionnel  $\mathbb{M}^Y$  qui contient les DGP pour les variables endogènes  $\mathbf{Y}_t$  conditionnellement aux variables explicatives  $\mathbf{X}_t$ ;
- (iii) le modèle complet  $\mathbb{M}$  comprend tous les DGP joints  $(\mu^Y, \mu^X)$ , où  $\mu^X$  est un élément arbitraire de  $\mathbb{M}^X$  et où  $\mu^Y$  est un élément arbitraire de  $\mathbb{M}^Y$ ; et
- (iv) il existe une application définissante des paramètres  $\theta^Y: \mathbb{M}^Y \rightarrow \Theta$  telle que, pour tout  $\mu \equiv (\mu^Y, \mu^X) \in \mathbb{M}$ ,  $\theta(\mu) = \theta^Y(\mu^Y)$ .

Cette définition nécessite quelques mots d'explication. Les DGP du sous-modèle  $\mathbb{M}^X$  sont caractérisés par des séries des contributions telles que  $\ell_t^X$  dans (18.11), alors que ceux de  $\mathbb{M}^Y$  sont caractérisés par des contributions telles que  $\ell_t^Y$  dans cette équation. Ainsi les contributions qui caractérisent les DGP des deux sous-modèles sont tels que, pour l'observation  $t$ , la densité est conditionnelle à *tous* les  $\Omega_t$ . Cela signifie en particulier que le processus qui génère les  $\mathbf{X}_t$  peut tout à fait dépendre des  $\mathbf{Y}_t$  retardés. La puissance de point (iii) de la définition est que le modèle complet  $\mathbb{M}$ , les DGP qui ont des contributions comparables au membre de droite de (18.11), doit contenir *toutes* les combinaisons d'éléments de  $\mathbb{M}^X$  et  $\mathbb{M}^Y$  possibles. Le point (iv) indique que les paramètres du modèle ne dépendent que du DGP conditionnel qui génère les  $\mathbf{Y}_t$  conditionnellement aux  $\mathbf{X}_t$ . Autrement dit, les paramètres associés au DGP  $(\mu^Y, \mu^X)$  ne dépendent que de  $\mu^Y$ . Si on remplace  $\mu^X$  par un autre DGP pour les mêmes variables explicatives, disons  $\nu^X$ , les paramètres ne sont pas modifiés.

Engle, Hendry, et Richard prétendent que l'exogénéité faible au sens de la définition précédente est précisément ce dont nous avons besoin pour estimer et réaliser des inférences sur les paramètres  $\theta$  without sans tenir compte du sous-modèle  $\mathbb{M}^X$ . Afin d'estimer les modèles par maximum de vraisemblance, cela est suffisamment clair. La fonction de log-vraisemblance est la somme des contributions du type (18.11). Seul le premier terme, issu du sous-modèle

$\mathbb{M}^Y$ , peut dépendre de  $\theta$ . La maximisation de la fonction de log-vraisemblance dans sa totalité est donc équivalente à la maximisation de la **fonction de log-vraisemblance partielle**

$$\ell^Y(\mathbf{Y}^n, \mathbf{X}^n; \theta) \equiv \sum_{t=1}^n \ell_t^Y(\mathbf{Y}_t | \mathbf{X}_t, \Omega_t; \theta)$$

par rapport à  $\theta$ . De la même façon, en ce qui concerne l'inférence, le gradient et la matrice Hessienne de la fonction de log-vraisemblance complète  $\ell$  par rapport à  $\theta$  sont identiques à ceux de la fonction de log-vraisemblance partielle  $\ell^Y$ .

Voyons comment s'applique la Définition 18.3 au modèle défini par (18.05) et (18.06). A l'évidence, (18.06) correspond au sous-modèle  $\mathbb{M}^X$  et (18.05) correspond au sous-modèle  $\mathbb{M}^Y$ . Notons que (18.06) fait usage des valeurs retardées de  $y_t$ . Remarquons que si les "paramètres"  $\delta_1$  et  $\delta_2$  étaient définis par l'application définissante des paramètres, l'exogénéité faible serait sans pertinence, puisque les  $\delta_i$  apparaissent *seulement* dans le sous-modèle  $\mathbb{M}^X$ . Pour éviter cette difficulté apparente, nous supposons que l'application définissante des paramètres ne définit que le paramètre  $\beta$ . Ainsi, dans ce cas, nous mettons les paramètres  $\delta_i$  et les éléments de la matrice de covariance  $\Sigma$  sur un pied d'égalité, en tant que paramètres perturbateurs. Le seul paramètre d'intérêt est  $\beta$ .

Un DGP du sous-modèle  $\mathbb{M}^X$  peut maintenant être spécifié en donnant les valeurs des paramètres perturbateurs  $\delta_i$  et la densité marginale des aléas  $\varepsilon_{2t}$ , qui dépendra de la variance non conditionnelle  $\sigma_{22}$  mais pas de  $\sigma_{11}$  ou de  $\sigma_{12}$ . Pour une DGP dans  $\mathbb{M}^Y$ , il est nécessaire de spécifier la valeur de  $\beta$ , le paramètre qui nous intéresse, et la densité de  $\varepsilon_{1t}$  conditionnellement à  $\varepsilon_{2t}$ , qui impliquera  $\sigma_{11}$  et  $\sigma_{12}$ . A ce stade, les conditions (i), (ii), et (iv) de la Définition 18.3 sont satisfaites. La variable  $x_t$  est donc faiblement exogène pour le modèle donné par (18.05), (18.06) et le paramètre  $\beta$  dès que la condition (iii) est satisfaite, ce qui implique que nous soyons capables d'associer deux DGP, *quels qu'ils soient*, correspondant chacun à un sous-modèle. Mais cela n'est pas possible en général, parce qu'il faut que  $\sigma_{11}\sigma_{22} \geq \sigma_{12}^2$  afin que la matrice de covariance de la distribution jointe de  $\varepsilon_{1t}$  et  $\varepsilon_{2t}$  soit semi-définie positive. Cette inégalité ne sera satisfaite automatiquement que si nous contraignons le modèle global de sorte que  $\sigma_{12} = 0$ , ce qui rend  $x_t$  faiblement exogène.

Nous voyons donc, dans ce cas, que la prédétermination de  $x_t$  se confond avec son exogénéité faible. Qu'advient-il si nous examinons le modèle donné par (18.08) et (18.06)? Souvenons-nous que  $x_t$  est prédéterminé dans (18.08) de manière tout à fait générale. En réalité, il sera également faiblement exogène en général si nous modifions l'application définissante des paramètres (mais pas le modèle  $\mathbb{M}$  sous-jacent) afin qu'elle décrive le paramètre  $b$  au lieu de  $\beta$ . Remarquons que même si nous nous intéressons aux paramètres  $c_1$ ,  $c_2$ ,

et à la variance des aléas  $v_t$  dans (18.08) autant qu'à  $b$ ,  $\beta$  ne peut pas être recomposé à partir de ces paramètres sans  $\sigma_{12}$ . L'exogénéité faible provient du fait que, par construction,  $v_t$  est non corrélé à  $\varepsilon_{2t}$ .

L'avantage de l'exogénéité faible par rapport à la prédétermination dans ce contexte est que sa définition fait référence à une application définissante des paramètres particulière. cela signifie que nous pouvons dire que  $x_t$  est faiblement exogène *pour*  $\beta$  ou pas, selon le cas, et qu'elle est toujours faiblement exogène *pour*  $b$ . A l'inverse, la prédétermination est définie relativement à une *équation*, telle que (18.05) ou (18.08), plutôt qu'à une application définissante des paramètres.

Le concept de **causalité au sens de Granger** est également un concept qui peut être important pour celui qui désire travailler conditionnellement à un ensemble de variables explicatives. Comme son nom le suggère, ce concept a été développé par Granger (1969). D'autres définitions de la causalité ont été proposées, en particulier par Sims (1972). Les définitions de la causalité au sens de Granger ou de Sims sont souvent équivalentes, mais pas toujours; consulter Chamberlain (1982) et Florens et Mouchart (1982). Pour la plupart des usages, il semble que la causalité au sens de Granger, ou plutôt son opposé, la **non causalité au sens de Granger**, soit le concept le plus utile.

Nous donnons à présent une définition de la non causalité au sens de Granger. Tout comme la définition de l'exogénéité faible, elle est relative au contexte des modèles  $\mathbb{M}$  qui contiennent les DGP qui génèrent deux ensembles de variables  $\mathbf{Y}_t$  et  $\mathbf{X}_t$ . Contrairement à celle-ci, elle ne fait référence à aucune application définissante des paramètres, et n'opère pas de distinction entre les variables endogènes  $\mathbf{Y}_t$  et les variables explicatives  $\mathbf{X}_t$ . Dans la définition,  $\mathbf{Y}^{t-1}$  et  $\mathbf{X}^{t-1}$  désignent les lignes des matrices  $\mathbf{Y}$  et  $\mathbf{X}$ , respectivement, antérieures à la  $t^{\text{th}}$ . Ainsi  $\Omega_t$  est composé de  $\mathbf{Y}^{t-1}$  et  $\mathbf{X}^{t-1}$ .

*Définition 18.4.*

Les variables  $\mathbf{Y}^{t-1}$  ne causent pas au sens de Granger les variables  $\mathbf{X}_t$  dans un modèle  $\mathbb{M}$  comprenant les DGP caractérisés par les contributions (18.11) si et seulement si

$$\ell_t^X(\mathbf{X}_t | \Omega_t) = \ell_t^X(\mathbf{X}_t | \mathbf{X}^{t-1}).$$

Cela signifie que  $\mathbf{Y}^{t-1}$  ne cause pas au sens de Granger  $\mathbf{X}_t$  si la distribution de  $\mathbf{X}_t$  conditionnellement au passé de  $\mathbf{X}_t$  et  $\mathbf{Y}_t$  est la même que celle qui est conditionnelle au passé de  $\mathbf{X}_t$ .

Un moyen pratique d'exprimer la non causalité au sens de Granger consiste à dire que le passé de  $\mathbf{Y}_t$  ne contient aucune information sur  $\mathbf{X}_t$  qui ne soit déjà contenue dans le passé de  $\mathbf{X}_t$ . Bien que cela ne soit pas strictement exact, il est fréquent de parler de causalité au sens de Granger plutôt que de non causalité au sens de Granger. Cette pratique n'entraîne en général aucune ambiguïté.

Il est évident à partir de (18.06) que, dans le modèle donné par cette équation et par (18.05),  $y_t$  *cause* au sens de Granger  $x_t$ , à moins que  $\delta_2 = 0$ . Ainsi, même si  $\sigma_{12} = 0$ , ce qui signifie que  $x_t$  est faiblement exogène pour le paramètre  $\beta$  dans (18.05), le processus générateur de  $x_t$  dépend du passé de la variable endogène  $y_t$ . par ailleurs, si  $\delta_2 = 0$  mais que  $\sigma_{12} \neq 0$ ,  $y_t$  ne cause pas  $x_t$  au sens de Granger, bien que  $x_t$  ne soit pas faiblement exogène pour  $\beta$ . Ainsi les deux idées de faible exogénéité et de non causalité au sens de Granger sont distinctes: aucune n'implique l'autre et aucune n'est impliquée par l'autre.

Comme nous l'avons vu, la présence de la causalité au sens de Granger ne nous empêche nullement d'estimer efficacement  $\beta$  et de réaliser des inférences sur ce paramètre sans avoir recours au processus qui génère  $x_t$  si  $x_t$  est faiblement exogène pour  $\beta$ . Inversement, une absence d'exogénéité faible ne nous empêche nullement de faire des *prévisions* efficaces de  $y_t$  conditionnellement à  $x_t$  si  $y_t$  ne cause pas  $x_t$  au sens de Granger. Plus précisément, supposons que nous établissions une équation d'anticipation de  $x_t$  basée sur son passé uniquement. Si (18.05) et (18.06) sont exactes, nous trouvons que

$$E(x_t | x^{t-1}) = (\delta_1 + \beta\delta_2)x_{t-1}. \quad (18.12)$$

On anticiperait alors  $x_t$  en termes de la valeur retardée  $x_{t-1}$  et d'une estimation du paramètre d'autorégression  $\delta_1 + \beta\delta_2$ , obtenu, sans doute, par une régression de  $x_t$  sur sa propre valeur retardée d'une période. Si par la suite nous souhaitons anticiper  $y_t$  conditionnellement à notre prévision de  $x_t$ , nous développerions une équation de prévision de  $y_t$  en fonction de celle de  $x_t$  et du passé des deux variables. De (18.08),

$$E(y_t | x_t, \Omega_t) = bx_t + c_1x_{t-1} + c_2y_{t-1}, \quad (18.13)$$

où  $b$ ,  $c_1$ , et  $c_2$  sont définis par (18.09). Si maintenant nous remplaçons  $x_t$  dans (18.13) par son anticipation (18.12), nous obtenons une prévision

$$b(\delta_1 + \beta\delta_2)x_{t-1} + c_1x_{t-1} + c_2y_{t-1}. \quad (18.14)$$

On déduit immédiatement de (18.05) et (18.06) que

$$E(y_t | \Omega_t) = \beta\delta_1x_{t-1} + \beta\delta_2y_{t-1}.$$

Par conséquent, si (18.14) doit procurer une anticipation sans biais, il est nécessaire que

$$b(\delta_1 + \beta\delta_2) + c_1 = \beta\delta_1 \quad \text{et} \quad c_2 = \beta\delta_2.$$

A l'aide des définitions (18.09), nous pouvons voir que ces égalités sont vérifiées si  $\delta_2 = 0$  ou si  $b = 0$ . La première condition est précisément celle de la non causalité au sens de Granger. La seconde correspond à un cas particulier où

$x_t$  ne contient aucune information sur  $y_t$  qui ne soit déjà contenue dans  $\Omega_t$ , et elle est moins intéressante dans le contexte actuel.

La conclusion en général est que lorsque nous portons notre attention sur la prévision, nous pouvons anticiper les valeurs des variables  $\mathbf{Y}_t$  conditionnellement aux anticipations sur les variables  $\mathbf{X}_t$  si  $\mathbf{Y}^{t-1}$  ne cause pas  $\mathbf{X}_t$  au sens de Granger. D'autre part, si nous portons notre attention sur l'estimation et l'inférence pour certains paramètres, nous pouvons conditionner par rapport à  $\mathbf{X}_t$  si ces variables sont faiblement exogènes pour les paramètres dans le contexte du modèle pour lequel ils sont définis. Il est intéressant de combiner les deux idées pour définir les circonstances pour lesquelles toutes des activités peuvent être entreprises avec succès conditionnellement à  $\mathbf{X}_t$ . Le concept approprié est celui de l'**exogénéité forte**, que nous définissons à présent.

*Définition 18.5.*

Les variables explicatives  $\mathbf{X}_t$  sont **fortement exogènes** pour le modèle paramétrisé  $(\mathbb{M}, \boldsymbol{\theta})$  comprenant les DGP qui génèrent à la fois les variables endogènes  $\mathbf{Y}_t$  et les  $\mathbf{X}_t$  si elles sont faiblement exogènes et si  $\mathbf{Y}^{t-1}$  ne cause pas  $\mathbf{X}_t$  au sens de Granger.

Ceci complète notre discussion sur la causalité et sur l'exogénéité. Pour une discussion encore plus complète, nous orientons les lecteurs vers l'article de Engle-Hendry-Richard. Au delà de l'introduction des concepts de faible et de forte exogénéité, cet article annonce un autre concept, appelé **super exogénéité**. Ce concept est important pour l'analyse politique, mais pas pour l'estimation ou l'inférence, et n'est donc pas dans notre priorité immédiate.

### 18.3 L'IDENTIFICATION DANS LES MODÈLES SIMULTANÉS

Le problème de l'identification dans les modèles d'équations simultanées est, en principe, comparable à ce dont nous avons discuté dans le contexte général des modèles paramétrisés. Si pour un modèle  $\mathbb{M}$  donné, il est possible de définir une application définissante des paramètres, alors les paramètres du modèle sont identifiés, dans le sens où un seul et unique vecteur de paramètres est associé à chaque DGP dans  $\mathbb{M}$ . Cependant, même si une telle application existe, les données doivent satisfaire certaines conditions pour que le modèle soit identifié par les données, et le DGP doit en satisfaire d'autres pour que le modèle soit identifié asymptotiquement. Dans le Chapitre 5, nous avons défini et discuté en détail du concept d'identification asymptotique, et nous l'avons comparé au concept d'identification par un ensemble d'observations particulier. Dans le cadre des modèles d'équations simultanées, c'est bien sûr le premier qui nous intéresse. Toutes les méthodes d'estimation que nous avons étudiées se fondent sur la théorie asymptotique, et on ne peut pas espérer réaliser des estimations convergentes si les paramètres ne sont pas identifiés asymptotiquement.

Dans cette section, nous traiterons de l'identification asymptotique d'un modèle d'équations simultanées par l'estimateur des doubles moindres carrés, que nous avons introduit dans la Section 7.5. Cela peut paraître un sujet limité, et dans un certains sens, c'est un sujet limité. Cependant, c'est un problème qui a donné naissance à une littérature très vaste, et que nous ne pouvons pas exposer en entier ici; voir Fisher (1976) et Hsiao (1983). Il existe des modèles qui ne sont pas identifiés par l'estimateur des 2SLS mais qui le sont par des d'autres, tels que l'estimateur FIML, et nous en parlerons brièvement. Il n'est pas très facile d'étendre la théorie que nous présentons dans le contexte des modèles non linéaires, contexte pour lequel il est habituellement recommandé de se recommander de se référer à la théorie asymptotique développée dans la Section 5.2.

Nous débutons par le modèle d'équations simultanées (18.01). Ce modèle comprend les DGP qui génèrent les échantillons d'où sont issus le vecteur  $\mathbf{Y}_t$  des  $g$  variables dépendantes, conditionnellement à un ensemble de variables exogènes et dépendante retardées  $\mathbf{X}_t$ . Puisque nous avons supposé que les variables exogènes  $\mathbf{X}_t$  sont faiblement exogènes, nous pouvons faire abstraction du processus qui les génère. Afin de poursuivre notre discussion sur l'identification, il faut poser quelques hypothèses sur les aléas  $\mathbf{U}_t$ . Il faut bien évidemment que  $E(\mathbf{U}_t) = \mathbf{0}$ , et il semble raisonnable de supposer qu'ils sont indépendants en série et que  $E(\mathbf{U}_t^\top \mathbf{U}_t) = \boldsymbol{\Sigma}_t$ , où  $\boldsymbol{\Sigma}_t$  est une matrice définie positive pour tout  $t$ . Si l'on veut réaliser de inférences à partir de la matrice de covariance des 2SLS, il est nécessaire d'imposer l'homoscédasticité des aléas, c'est-à-dire d'imposer  $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}$  pour tout  $t$ .

Il est pratique de traiter l'identification des paramètres équation par équation dans un modèle d'équations simultanées, puisqu'il est parfaitement envisageable d'identifier les paramètres d'une équation quelconque même si ceux des autre équations ne le sont pas. Pour simplifier la notation, nous ne considérerons, sans perte de généralité, que les paramètres de la première équation du système, c'est-à-dire les éléments des premières colonnes des matrices  $\boldsymbol{\Gamma}$  et  $\mathbf{B}$ . Comme nous l'avons noté dans la Section 18.1, il faut imposer des contraintes sur les éléments de ces matrices pour les identifier. Il est habituel de supposer que ces contraintes prennent toutes la forme de contraintes de nullité de certains paramètres. On dit qu'une variable est **exclue** d'une équation lorsque le coefficient correspondant est contraint à zéro; autrement, on parle de variable **incluse** dans l'équation. Comme nous l'avons vu dans la Section 6.4, il est toujours possible de reparamétriser les contraintes dans un contexte d'équation unique pour leur donner la forme de contraintes de nullité. Mais dans un contexte d'équations simultanées, de telle reparamétrisations n'existent en général qu'en l'absence de **contraintes d'équations croisées**, c'est-à-dire des contraintes qui impliquent les paramètres de plus d'une équation du système. S'il existe des contraintes d'équations croisées, alors il faut abandonner le contexte des systèmes linéaires, quoi que



l'on veuille tenter. Il nous faut également abandonner l'estimateur 2SLS si nous voulons imposer des contraintes d'équations croisées.

Partitionnons la matrice  $\mathbf{Y}$  comme suit:

$$\mathbf{Y} = [\mathbf{y} \quad \mathbf{Y}_1 \quad \mathbf{Y}_2], \quad (18.15)$$

où le vecteur colonne  $\mathbf{y}$  est la variable endogène associée au coefficient unitaire dans la première équation du système, les colonnes de la matrice  $\mathbf{Y}_1$  de dimension  $n \times g_1$  sont les variables endogènes *non* exclues de cette équation par des contraintes de nullité, et où les colonnes de la matrice  $\mathbf{Y}_2$  de dimension  $n \times (g - g_1 - 1)$  sont les variables endogènes exclues. Pareillement, nous partitionnons la matrice  $\mathbf{X}$  des variables exogènes:

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2], \quad (18.16)$$

où les colonnes de la matrice  $\mathbf{X}_1$  de dimension  $n \times k_1$  sont les variables exogènes qui sont incluses dans l'équation, et où celles de la matrice  $\mathbf{X}_2$  de dimension  $n \times (k - k_1)$  sont les variables exogènes exclues.

De façon cohérente avec la partition de  $\mathbf{Y}$  et  $\mathbf{X}$ , nous pouvons partitionner les matrices de coefficients  $\mathbf{\Gamma}$  et  $\mathbf{B}$  comme suit:

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & \mathbf{\Gamma}_{02} \\ -\gamma_1 & \mathbf{\Gamma}_{12} \\ \mathbf{0} & \mathbf{\Gamma}_{22} \end{bmatrix} \quad \text{et} \quad \mathbf{B} = \begin{bmatrix} \beta_1 & \mathbf{B}_{12} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix}. \quad (18.17)$$

Les lignes de  $\mathbf{\Gamma}$  sont partitionnées comme les colonnes de  $\mathbf{Y}$  dans (18.15), et celle de  $\mathbf{B}$  le sont comme les colonnes de  $\mathbf{X}$  dans (18.16). En plus de cela, nous avons partitionné les colonnes de  $\mathbf{\Gamma}$  et  $\mathbf{B}$  pour qu'elles puissent séparer les premières colonnes de chaque matrice des autres colonnes, puisque ce sont les premières colonnes qui contiennent les paramètres de la première équation du système. On peut donc écrire la première équation comme suit:

$$\mathbf{y} = \mathbf{Y}_1 \gamma_1 + \mathbf{X}_1 \beta_1 + \mathbf{u} = \mathbf{Z} \boldsymbol{\delta} + \mathbf{u}, \quad (18.18)$$

où la matrice  $\mathbf{Z}$  de dimension  $n \times (g_1 + k_1)$  est  $[\mathbf{X}_1 \quad \mathbf{Y}_1]$ , et où le vecteur paramétrique  $\boldsymbol{\delta}$  est  $[\beta_1 : \gamma_1]$ .

Pour obtenir une estimation 2SLS de  $\boldsymbol{\delta}$ , nous devons utiliser des variables instrumentales. Les colonnes de  $\mathbf{X}_1$ , qui sont exogènes, peuvent servir en tant qu'instruments, et celles de  $\mathbf{X}_2$  constituent des instruments supplémentaires. Si les colonnes de  $\mathbf{X}$  sont les seuls instruments disponibles, il va de soi qu'une condition nécessaire à l'identification de  $\boldsymbol{\delta}$ , que ce soit avec des échantillons finis ou asymptotiquement, est que  $\mathbf{X}$  possède au moins autant de colonnes que  $\mathbf{Z}$ . Cela revient à dire que  $\mathbf{X}_2$  doit posséder au moins autant de colonnes que  $\mathbf{Y}_1$ , c'est-à-dire que  $k - k_1 \geq g_1$ . Autrement dit, il faut que le nombre des variables exogènes exclues soit au moins aussi grand que celui des variables endogènes incluses. Cette condition est connue sous le nom de **condition d'ordre** pour l'identification. Cependant, comme nous le verrons, c'est une condition nécessaire mais qui n'est pas suffisante en général.<sup>1</sup>

<sup>1</sup> Si on admet la possibilité de contraintes d'équations croisées, cette condition d'ordre n'est plus du tout nécessaire.

Il n'est pas évident que  $\mathbf{X}$  fournisse toutes les variables instrumentales requises. Pourquoi ne pas employer d'autres variables endogènes ou prédéterminées qui sont corrélées aux variables endogènes  $\mathbf{Y}_1$ ? Même dans le cas où la condition d'ordre est vérifiée, ne pourrions-nous pas faire usage d'autres instruments disponibles pour obtenir des estimations plus efficaces? Il s'avère que l'usage d'instruments supplémentaires ne permet pas d'identifier asymptotiquement des paramètres qui ne le sont pas. De plus, lorsque les aléas  $\mathbf{u}$  sont homoscédastiques et indépendants en série, les instruments supplémentaires n'apportent aucun gain d'efficacité.

Pour mettre en évidence ces résultats, nous considérons la forme réduite contraintes (18.02) correspondant à (18.01). Par un léger abus de notation, nous poserons simplement

$$\mathbf{Y} = \mathbf{X}\Pi + \mathbf{V}, \quad (18.19)$$

en définissant  $\Pi$  par  $\mathbf{B}\Gamma^{-1}$ . Il sera nécessaire de partitionner  $\Pi$  conformément aux partitions (18.17) de  $\Gamma$  et  $\mathbf{B}$ :

$$\Pi = \begin{bmatrix} \pi_1 & \Pi_{11} & \Pi_{12} \\ \pi_2 & \Pi_{21} & \Pi_{22} \end{bmatrix}. \quad (18.20)$$

La partition des lignes est ici la même que celle de  $\mathbf{B}$  dans (18.17), et la partition des colonnes est identique à celle de  $\Gamma$  dans la même équation, ainsi qu'à celle de  $\mathbf{Y}$  dans (18.15). Nous supposons que les données ont été générées par le processus (18.19) avec  $\Pi = \Pi_0 = \mathbf{B}_0\Gamma_0^{-1}$ .

Considérons à présent l'identification du vecteur paramétrique  $\delta$  dans l'équation (18.18) pour n'importe quelle matrice  $\mathbf{W}$  d'instruments valables, c'est-à-dire n'importe quelle matrice  $\mathbf{W}$  telle que  $\text{plim}(n^{-1}\mathbf{W}^\top\mathbf{W})$  est une matrice définie et déterministe, et telle que  $\text{plim}(n^{-1}\mathbf{W}^\top\mathbf{V}) = \mathbf{0}$ . À partir des résultats de la Section 7.8,  $\delta$  est identifiable par les données si la matrice  $\mathbf{Z}^\top\mathbf{P}_W\mathbf{Z}$  est définie positive, et il est identifiable asymptotiquement si  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_W\mathbf{Z})$  est définie positive. Pour étudier cette limite en probabilité, examinons la matrice

$$\begin{aligned} \frac{1}{n}\mathbf{W}^\top\mathbf{Z} &= \frac{1}{n}\mathbf{W}^\top[\mathbf{X}_1 \quad \mathbf{Y}_1] \\ &= \frac{1}{n}\mathbf{W}^\top[\mathbf{X}_1 \quad \mathbf{X}_1\Pi_{11} + \mathbf{X}_2\Pi_{21} + \mathbf{V}_1], \end{aligned} \quad (18.21)$$

où le bloc  $\mathbf{V}_1$  de la matrice d'aléas  $\mathbf{V}$  correspond au bloc  $\mathbf{Y}_1$  de  $\mathbf{Y}$  dans (18.15), et où les coefficients de la forme réduite sont évalués avec  $\Pi = \Pi_0$ .

L'orthogonalité asymptotique entre les instruments  $\mathbf{W}$  et la matrice d'aléas  $\mathbf{V}$  signifie que la limite en probabilité de (18.21) est

$$\text{plim}_{n \rightarrow \infty} \left( \frac{1}{n}\mathbf{W}^\top[\mathbf{X}_1 \quad \mathbf{X}_1\Pi_{11} + \mathbf{X}_2\Pi_{21}] \right). \quad (18.22)$$

Ceci montre clairement que, quel que soit le choix d'une matrice d'instruments  $\mathbf{W}$ , le rang de la matrice (18.22) ne peut excéder  $k$ , qui est précisément le nombre de variables exogènes linéairement indépendantes. Toutes les colonnes de

la matrice partitionnée dans (18.22) sont des colonnes de  $\mathbf{X}$  ou des combinaisons linéaires de ces colonnes. Il s'ensuit que le rang de  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_W\mathbf{Z})$  ne peut jamais dépasser  $k$  lui non plus. Ainsi, si  $\mathbf{Z}$  possède plus de  $k$  colonnes, ce qui implique une violation de la condition d'ordre,  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_W\mathbf{Z})$  est singulière, et donc, non définie positive. Nous concluons que la condition d'ordre est bien nécessaire pour l'identification asymptotique de  $\boldsymbol{\delta}$ , quel que soit l'ensemble d'instruments employé.

Puis nous montrons que, sous les hypothèses d'homoscédasticité et d'indépendance en série des aléas  $\mathbf{u}$ , les colonnes de  $\mathbf{X}$  offrent des instruments optimaux pour l'estimation de  $\boldsymbol{\delta}$ . Il y a deux éventualités possibles. Dans la première,  $\mathcal{S}(\mathbf{X}) \subset \mathcal{S}(\mathbf{W})$ . Puisque  $\mathbf{X}_1$  et  $\mathbf{X}_2$  appartiennent à  $\mathcal{S}(\mathbf{X})$ , nous voyons à partir de (18.22) que

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{P}_W \mathbf{Z} \right) &= \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{Z}^\top \mathbf{P}_X \mathbf{Z} \right) \\ &= \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} [\mathbf{X}_1 \quad \mathbf{X}_1 \boldsymbol{\Pi}_{11} + \mathbf{X}_2 \boldsymbol{\Pi}_{21}]^\top [\mathbf{X}_1 \quad \mathbf{X}_1 \boldsymbol{\Pi}_{11} + \mathbf{X}_2 \boldsymbol{\Pi}_{21}] \right). \end{aligned}$$

Ainsi l'ajout d'instruments  $\mathbf{W}$  à ceux offerts par  $\mathbf{X}$  ne produit aucun gain d'efficacité asymptotique. Puisque cela contribuera à accroître le biais dans les échantillons finis (voir la Section 7.5), il vaut mieux ne pas utiliser ces instruments supplémentaires.

Dans la seconde,  $\mathcal{S}(\mathbf{X})$  n'est pas un sous-espace de  $\mathcal{S}(\mathbf{W})$ . Cela implique que, asymptotiquement,  $\mathbf{W}$  doit avoir un pouvoir explicatif sur  $\mathbf{Z}$  inférieur à celui de  $\mathbf{X}$ . Par conséquent,  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_X\mathbf{Z}) - \text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_W\mathbf{Z})$  est une matrice semi-définie positive pour toute matrice d'instruments  $\mathbf{W}$ . Il s'ensuit que (voir l'Annexe A)  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_W\mathbf{Z})^{-1} - \text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_X\mathbf{Z})^{-1}$  est également une matrice semi-définie positive. Ainsi la matrice de covariance asymptotique que l'on obtient à l'aide de la matrice d'instruments  $\mathbf{X}$ , à savoir  $\sigma^2 \text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_X\mathbf{Z})^{-1}$ , établit une borne inférieure pour la matrice de covariance asymptotique pour tout estimateur IV.

De la discussion précédente et des résultats de la Section 7.8, il ressort que la condition nécessaire et suffisante pour l'identification asymptotique de  $\boldsymbol{\delta}$  à l'aide des instruments optimaux  $\mathbf{X}$  est simplement que  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_X\mathbf{Z})$  soit non singulière. La littérature traditionnelle sur les modèles d'équations simultanées fait référence à cette condition en tant que **condition de rang** pour l'identification, pour des raisons évidentes. Cependant, un exposé aussi simple de cette condition est très rare. Au lieu de cela, la condition est typiquement exprimée en termes des coefficients de  $\boldsymbol{\Gamma}$  et  $\mathbf{B}$  de la forme structurelle ou des coefficients de la forme réduite contrainte. Etant donné que nous avons défini  $\boldsymbol{\Pi}$  en termes de  $\boldsymbol{\Gamma}$  et  $\mathbf{B}$  uniquement, toutes condition que l'on peut exprimer en termes d'un ensemble de coefficients peut s'exprimer en termes de l'autre.

Nous allons à présent montrer comment on peut exprimer la condition, qui veut que  $\text{plim}(n^{-1}\mathbf{Z}^\top\mathbf{P}_X\mathbf{Z})$  soit non singulière, en termes de contraintes

sur  $\Pi$  dans le DGP. Les paramètres  $\gamma_1$  et  $\beta_1$  de la première équation structurelle peuvent être identifiés si et seulement on peut les retrouver de façon unique à partir de la matrice  $\Pi$  des paramètres de la forme réduite contrainte. Cette matrice, par définition, satisfait l'équation  $\Pi\Gamma = B$ , dont nous pouvons écrire la première colonne sous la forme

$$\begin{aligned}\pi_1 + \Pi_{11}\gamma_1 &= \beta_1 \\ \pi_2 + \Pi_{21}\gamma_1 &= 0\end{aligned}$$

en vertu des partitions de (18.17) et (18.20). La première de ces deux équations sert à définir  $\beta_1$  en termes de  $\Pi$  et  $\gamma_1$ , et nous permet de voir que  $\beta_1$  peut être identifié si  $\gamma_1$  l'est aussi. La seconde équation montre que  $\gamma_1$  est déterminé de façon unique si et seulement si la sous-matrice  $\Pi_{21}$  est de plein rang en colonnes, c'est-à-dire si le rang de la matrice est égal au nombre de ses colonnes (voir l'Annexe A). La sous-matrice  $\Pi_{21}$  possède  $k - k_1$  lignes et  $g_1$  colonnes. Par conséquent, si la condition d'ordre est satisfaite, il y a au moins autant de lignes que de colonnes. La condition à l'identification de  $\gamma_1$ , mais aussi à celle de  $\beta_1$ , est que les colonnes de  $\Pi_{21}$  soient linéairement indépendantes.

Il est instructif de voir pourquoi cette dernière condition est équivalente à la condition de rang en termes de  $\text{plim}(n^{-1}Z^\top P_X Z)$ . Si, comme nous l'avons supposé tacitement tout au long de cette discussion, les variables exogènes  $X$  satisfont la condition que  $\text{plim}(n^{-1}X^\top X)$  est définie positive, alors  $\text{plim}(n^{-1}Z^\top P_X Z)$  peut ne pas être de plein rang si  $\text{plim}(n^{-1}X^\top Z)$  a un rang inférieur à  $g_1 + k_1$ , le nombre de colonnes de  $Z$ . La limite en probabilité de la matrice  $n^{-1}X^\top Z$  provient de (18.22), en remplaçant  $W$  par  $X$ . Si nous faisons abstraitin de la limite en probabilité et du facteur  $n^{-1}$  pour simplifier la notation, la matrice pertinente peut s'écrire comme suit:

$$\begin{bmatrix} X_1^\top X_1 & X_1^\top X_1 \Pi_{11} + X_1^\top X_2 \Pi_{21} \\ X_2^\top X_1 & X_2^\top X_1 \Pi_{11} + X_2^\top X_2 \Pi_{21} \end{bmatrix}. \quad (18.23)$$

La matrice (18.23) n'est pas de plein rang  $g_1 + k_1$  si et seulement s'il existe un vecteur non nul  $\theta \equiv [\theta_1 : \theta_2]$  de dimension  $(g_1 + k_1)$  tel que (18.23) fois ce vecteur donne un vecteur nul. Si nous explicitons cette condition, et si nous arrangeons les différents termes, nous obtenons

$$\begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} \begin{bmatrix} \theta_1 + \Pi_{11}\theta_2 \\ \Pi_{21}\theta_2 \end{bmatrix} = 0. \quad (18.24)$$

La première matrice du membre de gauche est simplement  $X^\top X$ , et elle est clairement non singulière. La condition porte alors sur les deux équations vectorielles

$$\theta_1 + \Pi_{11}\theta_2 = 0 \quad (18.25)$$

$$\Pi_{21}\theta_2 = 0. \quad (18.26)$$

Si ces équations sont vérifiées pour un vecteur  $\theta$  non nul, il est clair que  $\theta_2$  ne peut pas être nul. Par conséquent, la seconde équation n'est vérifiée que si  $\Pi_{21}$  n'est pas de plein rang. Alors si la condition de rang en termes de  $Z^T P_X Z$  n'est pas vérifiée, alors elle ne l'est pas non plus en termes de  $\Pi_{21}$ . Inversement, supposons que (18.26) soit vérifiée pour un vecteur  $\theta_2$  non nul quelconque de dimension  $g_1$ . Alors  $\Pi_{21}$  n'est pas de plein rang. Définissons  $\theta_1$  en termes de  $\theta_2$  et  $\Pi$  grâce à (18.25). Alors (18.25) et (18.26) impliquent ensemble (18.24), et la condition de rang initiale n'est pas satisfaite. Ainsi les deux versions de la condition de rang sont équivalentes.

Nous terminons cette section en établissant, sans démonstration, une troisième version de la condition de rang, équivalente aux deux premières, en termes des paramètres structurels  $\Gamma$  et  $B$ . Il est impossible d'exprimer cette condition exclusivement en termes des paramètres  $\gamma_1$  et  $\beta_1$  de la première équation. Au contraire, ce sont *uniquement* les valeurs des autres paramètres qui déterminent la possible identification de  $\gamma_1$  et  $\beta_1$ . Ce troisième exposé de la condition de rang est formulé de la manière suivante. Construisons la matrice de dimension  $(g - g_1 - 1 + k - k_1) \times (g - 1)$

$$\begin{bmatrix} \Gamma_{22} \\ B_{22} \end{bmatrix}.$$

Alors la condition de rang est satisfaite si et seulement si cette matrice est de plein rang  $g - 1$ .

Nous n'avons discuté dans cette section que des conclusions les plus importantes d'un programme de recherche ambitieux. Hsiao (1983) donne un traitement plus précis. Nous n'avons pas géré des problèmes tels que les contraintes d'équations croisées ou les contraintes impliquant la matrice de covariance  $\Sigma$ ; voir Rothenberg (1971), Richmond (1974), et Hausman et Taylor (1983), parmi d'autres. Dans la pratique, la condition d'ordre pour l'identification est beaucoup plus utile que la condition de rang parce qu'elle est beaucoup plus difficile à vérifier. Cependant, la condition de rang a un intérêt théorique certain, et il est instructif de voir qu'elle peut s'exprimer comme une condition très simple portant sur la limite en probabilité d'une certaine matrice qui doit être de plein rang. Elle est donc équivalente à la condition portant sur un certain estimateur 2SLS, celui qui utilise en tant qu'instruments toutes les variables exogènes et prédéterminées, qui doit avoir une matrice de covariance asymptotique non singulière.

## 18.4 MAXIMUM DE VRAISEMBLANCE EN INFORMATION COMPLÈTE

Il est possible d'établir une classification de deux façons des modèles d'équations simultanées. La première classification naturelle distingue les méthodes équation par équation des méthodes systémiques. Les premières, dont les représentants principaux sont les 2SLS et le LIML, estiment le modèle

équation par équation. Les secondes, dont les représentants principaux sont les 3SLS et le FIML, estiment tous les paramètres du modèle en même temps. Les adjectifs “information limitée” et “information complète” qui composent les noms LIML et FIML montrent clairement que la première méthode s’applique équation par équation, et que la seconde s’applique au système dans sa globalité. Les méthodes équation par équation sont plus faciles à mettre en oeuvre, alors que les méthodes systématiques produisent des estimations potentiellement plus efficaces.

L’autre classification naturelle distingue les méthodes basées sur le maximum de vraisemblance, à savoir le LIML et FIML, des méthodes basées sur les variables instrumentales ou la méthode des moments généralisés, dont les représentants les plus connus sont les 2SLS et les 3SLS. Les méthodes du ML produisent des estimations invariantes à la reparamétrisation (voir la Section 8.3) alors que ce n’est pas le cas des méthodes des IV. Nous avons déjà vu en détail les 2SLS dans le Chapitre 7. Au cours de cette section, nous fournirons un traitement détaillé de FIML, qui diffère des 2SLS quelle que soit la classification retenue. Les sections suivantes seront consacrées au LIML et aux 3SLS.

Tous les estimateurs d’équations simultanées tentent de gérer le fait que les aléas des équations structurelles sont corrélés avec n’importe quelle variable endogène apparaissant dans l’équation. Cette corrélation rend les OLS non convergents. Nous avons vu que les 2SLS gèrent ce problème en remplaçant les régresseurs défectueux par des instruments. D’un autre côté, le FIML gère ce problème par la maximisation d’une fonction de log-vraisemblance qui implique un terme Jacobien qui n’est pas simplement la transformation d’une somme de résidus au carré. Le FIML gère également deux problèmes qui se manifestent dans le cadre de tout modèle multivarié, qu’il y ait ou non simultanété; voir la Section 9.9. Le premier problème est que, en dehors de rares cas, les aléas des différentes équations seront corrélés. Les techniques équation par équation telles que les 2SLS ou le LIML ignorent purement et simplement ce problème. Au contraire, les techniques systématiques telles que le FIML ou les 3SLS assurent la gestion de ce problème et devraient normalement produire des estimations plus efficaces en général. Le second problème est que, dans de nombreux modèles, il existe des contraintes d’équations croisées. Les méthodes équation par équation ignorent nécessairement ce problème, mais les méthodes systématiques telles que le FIML en tiennent compte. Lorsque le système complet est établi, les paramètres qui apparaissent dans plus d’une équation sont automatiquement traités de façon différente des paramètres qui n’apparaissent que dans une seule.

Le modèle d’équations simultanées linéaire (18.01), dont les aléas sont supposés être normalement distribués, homoscédastiques et indépendants en série, peut s’écrire

$$\mathbf{Y}_t \boldsymbol{\Gamma} = \mathbf{X}_t \mathbf{B} + \mathbf{U}_t, \quad \mathbf{U}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad (18.27)$$

avec une notation qui est désormais familière. Souvenons-nous simplement que  $\mathbf{Y}_t$  est de dimension  $1 \times g$ ,  $\mathbf{\Gamma}$  est de dimension  $g \times g$ ,  $\mathbf{X}_t$  est de dimension  $1 \times k$ ,  $\mathbf{B}$  est de dimension  $k \times g$ ,  $\mathbf{U}_t$  est de dimension  $1 \times g$ , et  $\mathbf{\Sigma}$  est de dimension  $g \times g$ . Le moyen le plus simple d'obtenir la densité de  $\mathbf{Y}_t$  consiste à écrire celle de  $\mathbf{U}_t$ :

$$(2\pi)^{-g/2} |\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{U}_t \mathbf{\Sigma}^{-1} \mathbf{U}_t^\top\right).$$

Puis nous remplaçons  $\mathbf{U}_t$  par  $\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B}$  et multiplions par un terme Jacobien approprié. ce terme est la valeur absolue du déterminant du Jacobien de la transformation de  $\mathbf{Y}_t$  en  $\mathbf{U}_t$ , c'est-à-dire le déterminant de  $\mathbf{\Gamma}$ . Ainsi le facteur Jacobien est  $|\det \mathbf{\Gamma}|$ .<sup>2</sup> Le résultat est

$$(2\pi)^{-g/2} |\det \mathbf{\Gamma}| |\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B}) \mathbf{\Sigma}^{-1} (\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B})^\top\right).$$

De là, nous voyons que la fonction de log-vraisemblance est

$$\begin{aligned} \ell(\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Sigma}) &= \sum_{t=1}^n \ell_t(\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Sigma}) = -\frac{ng}{2} \log(2\pi) + n \log |\det \mathbf{\Gamma}| \\ &\quad - \frac{n}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} \sum_{t=1}^n (\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B}) \mathbf{\Sigma}^{-1} (\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B})^\top. \end{aligned} \quad (18.28)$$

Une première étape pratique dans la maximisation de  $\ell(\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Sigma})$  consiste à la concentrer par rapport à  $\mathbf{\Sigma}$  ou, comme nous l'avons fait dans la Section 9.9, par rapport à son inverse,  $\mathbf{\Sigma}^{-1}$ . Etant donné que

$$\frac{\partial \ell}{\partial \mathbf{\Sigma}^{-1}} = \frac{n}{2} \mathbf{\Sigma} - \frac{1}{2} \sum_{t=1}^n (\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B})^\top (\mathbf{Y}_t \mathbf{\Gamma} - \mathbf{X}_t \mathbf{B}),$$

(voir Annexe A) il est évident que

$$\mathbf{\Sigma}(\mathbf{B}, \mathbf{\Gamma}) = \frac{1}{n} (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B})^\top (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B}). \quad (18.29)$$

Nous pouvons substituer (18.29) à  $\mathbf{\Sigma}$  dans (18.28) pour obtenir

$$\begin{aligned} \ell^c(\mathbf{B}, \mathbf{\Gamma}) &= -\frac{ng}{2} (\log(2\pi) + 1) + n \log |\det \mathbf{\Gamma}| \\ &\quad - \frac{n}{2} \log \left| \frac{1}{n} (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B})^\top (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B}) \right|. \end{aligned} \quad (18.30)$$

<sup>2</sup> Dans ce chapitre, nous notons  $|\mathbf{A}|$  le déterminant de  $\mathbf{A}$  et  $|\det \mathbf{A}|$  la valeur absolue du déterminant. il est nécessaire d'employer la notation "det", que nous préférons éviter par ailleurs, lorsque la valeur absolue apparaît dans la formule.

Cette fonction de log-vraisemblance concentrée ressemble étroitement à (9.65), la fonction de log-vraisemblance concentrée pour un modèle de régression multivariée. Remarquons que nous avons usé de la même astuce que pour évaluer le second terme de la dernière ligne de (18.28). La différence entre (9.65) et (18.30) provient de la présence du terme Jacobien  $n \log |\det \mathbf{\Gamma}|$ , dont nous allons évaluer le rôle plus tard. L'estimateur FIML ne sera pas défini si la matrice  $(\mathbf{Y}\mathbf{\Gamma} - \mathbf{X}\mathbf{B})^\top(\mathbf{Y}\mathbf{\Gamma} - \mathbf{X}\mathbf{B})$  qui apparaît dans (18.30) n'est pas de plein rang pour toutes les valeurs admissibles de  $\mathbf{B}$  et  $\mathbf{\Gamma}$ , et cela nécessite que  $n \geq g + k$ . Ce résultat suggère également que  $n$  doit être suffisamment grand par rapport à  $g + k$  pour conserver au FIML de bonnes propriétés; consulter Sargan (1975) et Brown (1981).

Il est révélateur de dériver cette fonction de log-vraisemblance concentrée d'une manière radicalement opposée. Cette fois, nous partons de la forme réduire contrainte correspondant à (18.27), qui est

$$\mathbf{Y}_t = \mathbf{X}_t \mathbf{B} \mathbf{\Gamma}^{-1} + \mathbf{V}_t. \quad (18.31)$$

Ce système d'équations est juste un cas particulier du modèle de régression multivariée étudié dans la Section 9.9, mais sous la forme (9.43), avec un ensemble de fonctions de régression donné par  $\boldsymbol{\xi}_t \equiv \mathbf{X}_t \mathbf{B} \mathbf{\Gamma}^{-1}$  et qui sont des fonctions non linéaires des éléments de  $\mathbf{B}$  et  $\mathbf{\Gamma}$ . La fonction de log-vraisemblance concentrée correspondant à (18.31) est par conséquent (9.65). dans notre cas particulier, (9.65) devient

$$-\frac{ng}{2}(\log(2\pi) + 1) - \frac{n}{2} \log \left| \frac{1}{n} (\mathbf{Y} - \mathbf{X} \mathbf{B} \mathbf{\Gamma}^{-1})^\top (\mathbf{Y} - \mathbf{X} \mathbf{B} \mathbf{\Gamma}^{-1}) \right|. \quad (18.32)$$

Cette nouvelle expression pour  $\ell^c(\mathbf{B}, \mathbf{\Gamma})$  est égale à celle dérivée précédemment, (18.30). L'égalité entre (18.30) et (18.32) découle du fait que

$$\begin{aligned} & -\frac{n}{2} \log \left| \frac{1}{n} (\mathbf{Y} - \mathbf{X} \mathbf{B} \mathbf{\Gamma}^{-1})^\top (\mathbf{Y} - \mathbf{X} \mathbf{B} \mathbf{\Gamma}^{-1}) \right| \\ &= -\frac{n}{2} \log \left| \frac{1}{n} (\mathbf{\Gamma}^\top)^{-1} \mathbf{\Gamma}^\top (\mathbf{Y} - \mathbf{X} \mathbf{B} \mathbf{\Gamma}^{-1})^\top (\mathbf{Y} - \mathbf{X} \mathbf{B} \mathbf{\Gamma}^{-1}) \mathbf{\Gamma} \mathbf{\Gamma}^{-1} \right| \\ &= n \log |\det \mathbf{\Gamma}| - \frac{n}{2} \log \left| \frac{1}{n} (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B})^\top (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B}) \right|. \end{aligned}$$

Il est intéressant de noter que la fonction de log-vraisemblance concentrée pour un modèle d'équations simultanées peut s'écrire de deux manières différentes, (18.30) et (18.32). Cela montre de façon tout à fait claire que les formes structurelle et réduite contrainte sont simplement des moyens d'exprimer le même modèle. Nous pouvons assimiler le modèle d'équations simultanées soit à un type particulier de modèle, dont la fonction de log-vraisemblance concentrée est donnée par (18.30), soit à un cas particulier de modèle de régression multivariée *non linéaire*, dont la fonction de log-vraisemblance concentrée est identique à celle de n'importe quel autre modèle



de régression multivariée. Mis sous cette forme, nous pouvons lui appliquer tous les résultats déjà établis dans le Chapitre 9 pour les modèles de régression multivariée. Cependant, parce que la matrice des coefficients  $\mathbf{B}\mathbf{\Gamma}^{-1}$  dépend non linéairement des coefficients de toutes les équations du modèle, (18.32) est en général moins pratique que (18.30).

Lorsqu'il fut proposé à l'origine par les chercheurs de la Commission Cowles (Koopmans, 1950), le FIML n'était pas d'un calcul aisé, parce que la maximisation de la fonction de log-vraisemblance (18.30) nécessite une optimisation numérique. Au fur et à mesure que les ordinateurs devenaient plus puissants et que ce genre de calcul se démocratisait, un certain nombre de procédures de maximisation de la fonction de log-vraisemblance fut proposé, et la plupart des progiciels d'économétrie modernes incorpore au moins l'une d'elles. Rothenberg et Leenders (1964), Chow (1968), Hausman (1974, 1975), et Dagenais (1978) sont des références à consulter sur ce thème.

Comme d'habitude, la matrice de covariance asymptotique des estimations paramétriques FIML  $\hat{\mathbf{B}}$ ,  $\hat{\mathbf{\Gamma}}$ , et  $\hat{\mathbf{\Sigma}}$  peut être estimée de différentes façons. Une approche qui reste relativement aisée mais peu recommandée avec de petits échantillons consiste à exécuter une régression OPG. Cette régression artificielle peut se baser sur la fonction de log-vraisemblance concentrée (18.28), mais pas sur la fonction concentrée (18.30), parce que cette dernière n'est pas écrite sous la forme d'une somme de contributions. Une deuxième approche consiste à partir de la forme (18.32) de la fonction de log-vraisemblance. Comme nous l'avons mis en évidence dans la Section 9.9, le bloc de la matrice d'information associé aux paramètres des fonctions de régression d'un modèle de régression multivariée est donné par (9.69), et ce bloc peut s'obtenir à l'aide de la GNR (9.58). Une troisième approche pour estimer la matrice de covariance asymptotique de  $\hat{\mathbf{B}}$  et  $\hat{\mathbf{\Gamma}}$  consiste à utiliser la propriété d'équivalence asymptotique entre les 3SLS et le FIML; nous verrons cette approche dans la Section 18.6.

Le terme Jacobien  $\log|\det \mathbf{\Gamma}|$  qui apparaît explicitement dans (18.30) joue un rôle fondamental dans l'estimation. Sa présence est essentielle à la convergence des estimations ML. De plus, lorsque le déterminant de  $\mathbf{\Gamma}$  tend vers zéro, ce terme tend vers l'infini. Ainsi la fonction de log-vraisemblance doit tendre vers moins l'infini chaque fois que le déterminant de  $\mathbf{\Gamma}$  tend vers zéro. Cela est cohérent, parce que le modèle n'est pas gérable si  $|\det \mathbf{\Gamma}| = 0$ , ce qui implique que la vraisemblance d'un tel ensemble de paramètres est nul. De fait, cela signifie que l'espace des valeurs possibles de  $\mathbf{\Gamma}$  est divisé en un certain nombre de régions, séparées par des singularités lorsque  $|\det \mathbf{\Gamma}| = 0$ . Dans le cadre du modèle d'offre-demande discuté dans la Section 7.3, par exemple, il n'existe qu'une seule singularité, qui survient lorsque les pentes des fonctions d'offre et de demande sont égales. On ne peut pas espérer qu'un algorithme de maximisation numérique passe à travers ces singularités en général, même si cela peut arriver. Ainsi, lorsque nous tentons de maximiser numériquement une fonction de log-vraisemblance, il y a peu de chances que nous trouvions le

maximum global si la région dans laquelle l'algorithme débute ne le contient pas. Cela suggère qu'il peut être très important de bien choisir les valeurs initiales lorsque nous employons le FIML.

Bien que le FIML se base sur l'hypothèse que les aléas sont normaux multivariés, cette hypothèse n'est pas nécessaire pour que les estimations  $\hat{\mathbf{B}}$  et  $\hat{\mathbf{T}}$  soient convergentes et asymptotiquement normales. Lorsque le FIML est employé alors que les aléas ne sont pas normalement distribués, c'est davantage un estimateur QML qu'un estimateur ML, et il ne sera pas asymptotiquement efficace. Comme nous l'avons vu dans la Section 9.6, tout modèle de régression peut être estimé de façon satisfaisante par le ML sous l'hypothèse de distribution normale des aléas, que celle-ci soit exacte ou pas. Ce résultat s'applique aussi au FIML parce que, comme le montre (18.32), celui-ci estime en fait un certain modèle de régression multivariée non linéaire. Toutefois, lorsque le modèle d'équations simultanées sous-jacent est non linéaire, ce résultat ne s'applique plus automatiquement; voir Phillips (1982).

Les tests de spécification du modèle sont aussi importants pour les modèles d'équations simultanées que pour les autres modèles économétriques. Le large éventail des tests classiques — LM, LR, Wald, et  $C(\alpha)$  — est bien sûr disponible à cet égard. Cependant, du fait que l'estimation FIML est relativement coûteuse et difficile, les utilisateurs peuvent être tentés de renoncer à un programme de tests de spécification ambitieux pour les modèles estimés par FIML. Il est par conséquent utile de garder à l'esprit le fait que de nombreux types de mauvaise spécification du modèle structurel (18.01) impliquent une mauvaise spécification similaire de la forme réduite contrainte (18.03). Par exemple, si un aléa quelconque du modèle structurel était corrélé en série, alors, à de très rares exceptions près, tous les aléas de la forme réduite contrainte doivent l'être aussi. De manière comparable, si un aléa quelconque était hétéroscédastique, alors tous les aléas de la forme réduite doivent l'être. Pareillement, si les paramètres du modèle structurel sont non constants sur l'échantillon, les paramètres de la FRL ne seront pas constants non plus. Puisque les équations de la FRL sont estimées par moindres carrés ordinaires, il est très facile de les tester contre des mauvaises spécifications telles que la corrélation en série, l'hétéroscédasticité, ou encore la non constance des coefficients. Si de tels phénomènes sont mis en évidence par les tests, on peut raisonnablement conclure que le modèle structurel est mal spécifié, même s'il n'a pas encore été estimé. L'inverse n'est pas exact, cependant, puisque ces tests peuvent manquer de puissance, en particulier si une seule équation structurelle est mal spécifiée.

Un test de mauvaise spécification supplémentaire que l'on devrait toujours mener est celui des **contraintes de suridentification**. Dans la Section 7.8, nous avons examiné la manière de tester des contraintes de suridentification pour une équation unique estimée par IV ou 2SLS. Nous sommes à présent intéressés par toutes les contraintes de suridentification pour le système dans sa globalité. Le nombre des degrés de liberté pour le test est égal au nombre

d'éléments dans la matrice  $\mathbf{\Pi}$  de la FRL,  $gk$ , moins le nombre de paramètres libres de  $\mathbf{B}$  et  $\mathbf{\Gamma}$ . Dans la plupart des cas, il y aura quelques contraintes de suridentification, et dans de nombreux cas, il y en aura un grand nombre. La manière la plus naturelle de les tester est probablement d'employer un test LR. La valeur contrainte de la fonction de log-vraisemblance est la valeur de (18.30) évaluée avec les estimations FIML  $\hat{\mathbf{B}}$  et  $\hat{\mathbf{\Gamma}}$ , et la valeur non contrainte est

$$-\frac{ng}{2}(\log(2\pi) + 1) - \frac{n}{2} \log \left| \frac{1}{n} (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}}) \right|, \quad (18.33)$$

où  $\hat{\mathbf{\Pi}}$  désigne les estimations OLS des paramètres de la FRL. Comme d'habitude, le double de la différence entre les valeurs contrainte et non contrainte de la fonction de log-vraisemblance sera asymptotiquement distribuée suivant un  $\chi^2$  dont le nombre de degrés de liberté est égal à celui des contraintes de suridentification. Si l'on s'attend à ce que ces contraintes de suridentification soient enfreintes et si l'on ne veut pas s'embarrasser de l'estimation du modèle structurel, on peut employer un test de Wald, comme Byron (1974) l'a suggéré.

Nous n'avons pas encore expliqué pourquoi les estimations OLS  $\hat{\mathbf{\Pi}}$  sont également les estimations ML. On voit aisément à partir de (18.33) que, pour obtenir des estimations ML de  $\mathbf{\Pi}$ , il est nécessaire de minimiser le déterminant

$$|(\mathbf{Y} - \mathbf{X}\mathbf{\Pi})^\top (\mathbf{Y} - \mathbf{X}\mathbf{\Pi})|. \quad (18.34)$$

Supposons que l'on évalue ce déterminant avec un ensemble d'estimations  $\hat{\mathbf{\Pi}}$  quelconque différent de  $\hat{\mathbf{\Pi}}$ . Puisqu'il est toujours possible d'écrire  $\hat{\mathbf{\Pi}} = \hat{\mathbf{\Pi}} + \mathbf{A}$  pour une certaine matrice  $\mathbf{A}$ , (18.34) devient

$$\begin{aligned} & |(\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}} - \mathbf{X}\mathbf{A})^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{\Pi}} - \mathbf{X}\mathbf{A})| \\ &= |(\mathbf{M}_X \mathbf{Y} - \mathbf{X}\mathbf{A})^\top (\mathbf{M}_X \mathbf{Y} - \mathbf{X}\mathbf{A})| \\ &= |\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} + \mathbf{A}^\top \mathbf{X}^\top \mathbf{X} \mathbf{A}|. \end{aligned} \quad (18.35)$$

Parce que le déterminant de la somme de deux matrices définies positives est toujours supérieur à chacun des déterminants des deux matrices (voir l'Annexe A), il vient de (18.35) que (18.34) sera supérieur à  $\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}$  pour toute matrice  $\mathbf{A} \neq \mathbf{0}$ . Cela implique que  $\hat{\mathbf{\Pi}}$  minimise (18.34), ce qui démontre que les estimations OLS équations par équation de la FRL sont également les estimations ML systémiques.

Si l'on ne dispose pas d'un progiciel de régression qui calcule (18.33), il existe un moyen différent d'y parvenir. Considérons le **système récursif**

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}\boldsymbol{\eta}_1 + \mathbf{e}_1 \\ \mathbf{y}_2 &= \mathbf{X}\boldsymbol{\eta}_2 + \mathbf{y}_1\alpha_1 + \mathbf{e}_2 \\ \mathbf{y}_3 &= \mathbf{X}\boldsymbol{\eta}_3 + [\mathbf{y}_1 \quad \mathbf{y}_2]\alpha_2 + \mathbf{e}_3 \\ \mathbf{y}_4 &= \mathbf{X}\boldsymbol{\eta}_4 + [\mathbf{y}_1 \quad \mathbf{y}_2 \quad \mathbf{y}_3]\alpha_3 + \mathbf{e}_4, \end{aligned} \quad (18.36)$$

et ainsi de suite, où  $\mathbf{y}_i$  désigne la  $i^{\text{ième}}$  colonne de  $\mathbf{Y}$ . On peut interpréter ce système d'équations comme une simple reparamétrisation de la FRL (18.03). Il est aisé de voir que si l'on estime ces équations par OLS, tous les vecteurs de résidus seront orthogonaux:  $\hat{\mathbf{e}}_2$  sera orthogonal à  $\hat{\mathbf{e}}_1$ ,  $\hat{\mathbf{e}}_3$  sera orthogonal à  $\hat{\mathbf{e}}_2$  et à  $\hat{\mathbf{e}}_1$ , et ainsi de suite. Conformément à la FRL, tous les  $\mathbf{y}_i$  sont des combinaisons linéaires des colonnes de  $\mathbf{X}$  et d'erreurs aléatoires. Par conséquent, les équations de (18.36) sont correctes pour tout choix arbitraire des paramètres  $\alpha$ : les  $\boldsymbol{\eta}_i$  s'ajustent simplement selon le choix opéré. Toutefois, si nous *réclamons* l'orthogonalité des termes d'erreur  $\mathbf{e}_i$ , cela sert à identifier un choix particulier unique des  $\alpha$ . En réalité, le système récursif (18.36) possède autant de paramètres que la FRL (18.03):  $g$  vecteurs  $\boldsymbol{\eta}_i$ , possédant chacun  $k$  éléments,  $g - 1$  vecteurs  $\boldsymbol{\alpha}_i$ , avec en tout  $g(g - 1)/2$  paramètres, et  $g$  paramètres de variance, ce qui donne un total général de  $gk + (g^2 + g)/2$  paramètres. la FRL possède  $gk$  paramètres pour la matrice de covariance  $\boldsymbol{\Pi}$  et  $(g^2 + g)/2$  pour la matrice de covariance  $\boldsymbol{\Omega}$ , ce qui donne un total identique. La différence est que les paramètres  $\alpha$  de (18.36) ont été remplacés par les éléments non diagonaux de la matrice de covariance de  $\mathbf{V}$  dans la FRL.

Etant donné que le système récursif (18.36) est une simple reparamétrisation de la FRL (18.03), il ne devrait pas être surprenant d'apprendre que la fonction de log-vraisemblance pour le système récursif est égale à (18.33). Parce que les résidus des diverses équations dans (18.36) sont orthogonaux, la valeur des fonctions de log-vraisemblance des estimations OLS des équations individuelles. Ce résultat, que les lecteurs peuvent aisément vérifier numériquement, fournit parfois un moyen pratique de calculer la fonction de log-vraisemblance de la FRL. En dehors de cet usage, les systèmes récursifs sont d'une faible utilité. Ils ne procurent aucune information que ne soit déjà disponible dans la FRL, et la reparamétrisation dépend de l'ordonnancement des équations.

## 18.5 MAXIMUM DE VRAISEMBLANCE À INFORMATION LIMITÉE

L'un des problèmes qui se pose avec le FIML et les autres méthodes systémiques est qu'elles nécessitent de la part du chercheur une spécification de la structure de toutes les équations du modèle. La mauvaise spécification d'une équation quelconque conduira en général à des estimations non convergentes pour toutes les équations. Pour éviter ce problème, à condition que l'efficacité ne soit pas cruciale, les chercheurs peuvent préférer employer des méthodes équations par équation. La plus facile et la plus répandue est la méthode des 2SLS, mais elle souffre de deux inconvénients majeurs. les estimations qu'elle produit ne sont pas invariantes à la reparamétrisation, et, comme nous l'avons vu dans la Section 7.5, elles peuvent être sévèrement biaisées avec de petits échantillons. La méthode LIML est une technique alternative qui produit des estimations invariantes et qui, à de nombreux égards, possède de meilleures propriétés avec des échantillons finis que les 2SLS. Bien qu'elle ait été proposée

par Anderson et Rubin (1949) avant l'invention des 2SLS, et qu'elle ait été l'objet d'une étude plus théorique, elle a été peu utilisée par les économètres dans la pratique.

Comme son nom le suggère, l'idée de base du LIML consiste à employer une information partielle sur la structure du modèle. Supposons que l'on veuille estimer une seule équation, disons la première, d'un modèle structurel comme (18.01). Nous avons écrit une équation comparable dans la Section 18.3 sous la forme (18.18). Nous devons prendre en compte le fait que certaines variables apparaissant dans le membre de droite de (18.18), celles qui correspondent aux colonnes de  $\mathbf{Y}_1$ , sont endogènes. Le meilleur moyen d'en tenir compte consiste à écrire leurs équations sous la forme réduite libre:

$$\mathbf{Y}_1 = \mathbf{X}_1 \boldsymbol{\Pi}_{11} + \mathbf{X}_2 \boldsymbol{\Pi}_{21} + \mathbf{V}_1, \quad (18.37)$$

où la notation est identique à celle utilisée dans la Section 18.3. La combinaison de (18.18) et (18.37) donne le système d'équations

$$\begin{aligned} \mathbf{y} - \mathbf{Y}_1 \boldsymbol{\gamma}_1 &= \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{u} \\ \mathbf{Y}_1 &= \mathbf{X}_1 \boldsymbol{\Pi}_{11} + \mathbf{X}_2 \boldsymbol{\Pi}_{21} + \mathbf{V}_1. \end{aligned} \quad (18.38)$$

Remarquons que  $\mathbf{Y}_2$  n'apparaît plus du tout dans ce système d'équations. Si nous focalisons notre attention sur la première équation, les variables endogènes qui n'y apparaissent pas sont sans intérêt. On peut estimer le système d'équations (18.38) par maximum de vraisemblance, et les estimations  $\boldsymbol{\gamma}_1$  et  $\boldsymbol{\beta}_1$  qui en résultent seront les estimations LIML. Tout progiciel de FIML peut être employé à cette fin.

En fait, nous n'avons pas besoin d'un progiciel de FIML pour obtenir des estimations ML de (18.38). La matrice de coefficients des variables endogènes dans ce système d'équations est

$$\begin{bmatrix} 1 & \mathbf{0} \\ -\boldsymbol{\gamma}_1 & \mathbf{I} \end{bmatrix}. \quad (18.39)$$

Parce que cette matrice est triangulaire, son déterminant est simplement le produit des termes de la diagonale, et sa valeur est 1. Ainsi le terme Jacobien dans la fonction de log-vraisemblance disparaît, et la fonction de log-vraisemblance pour (18.38) a la même forme que celle de n'importe quel ensemble de régression apparemment sans lien (voir la Section 9.9). Cela implique que l'on peut utiliser n'importe quel programme pour l'estimation des systèmes SUR pour obtenir des estimations LIML. De plus, l'application des GLS faisables à un système tel que (18.38), en débutant par des estimations 2SLS pour la première équation et OLS pour les équations restantes, produira des estimations asymptotiquement équivalentes aux estimations LIML. Pagan (1979) a suggéré une procédure où l'on itère la procédure de GLS faisables jusqu'à ce qu'elle converge vers les véritables estimations LIML.

Dans la pratique, on calcule rarement les estimations LIML de cette façon, parce qu'il existe une méthode plus efficace pour les calculer. Il faudrait disposer de davantage d'outils algébriques pour la développer, mais les résultats terminaux seront relativement simples. À partir de (18.30), (18.32), et du fait que  $|\boldsymbol{\Gamma}| = 1$ , nous voyons que les estimations ML peuvent s'obtenir en minimisant

$$|(\mathbf{Y} - \mathbf{XB}\boldsymbol{\Gamma}^{-1})^\top(\mathbf{Y} - \mathbf{XB}\boldsymbol{\Gamma}^{-1})| = |(\mathbf{Y}\boldsymbol{\Gamma} - \mathbf{XB})^\top(\mathbf{Y}\boldsymbol{\Gamma} - \mathbf{XB})|. \quad (18.40)$$

Nous allons maintenant montrer que la minimisation du déterminant dans le membre de droite est ici équivalente à la minimisation du rapport de formes quadratiques, et que cela peut être réalisé, à son tour, en résolvant un certain problème de valeurs propres.

Ecrivons tout d'abord la matrice  $\mathbf{B}\boldsymbol{\Gamma}^{-1}$  qui apparaît dans le membre de gauche de (18.40). De (18.17) et d'une expression pour l'inverse de (18.39), nous voyons que

$$\mathbf{B}\boldsymbol{\Gamma}^{-1} = \begin{bmatrix} \beta_1 & \mathbf{B}_{12} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \gamma_1 & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \beta_1 + \mathbf{B}_{12}\gamma_1 & \mathbf{B}_{12} \\ \mathbf{B}_{22}\gamma_1 & \mathbf{B}_{22} \end{bmatrix}.$$

La matrice la plus à droite est simplement la version contrainte de  $\boldsymbol{\Pi}$ . L'élément au "nord-ouest" correspond à  $\mathbf{X}_1$  et la matrice au "sud-est" correspond à  $\mathbf{X}_2$ . Puisque  $\beta_1$  n'apparaît pas dans la matrice du bas et peut varier librement, il est clair que, quelle que soit la valeur de  $\gamma_1$ , nous pouvons trouver des valeurs de  $\beta_1$  et  $\mathbf{B}_{12}$  telles que l'élément au "nord-ouest" prenne n'importe quelle valeur. Autrement dit, les contraintes sur l'équation structurelle (18.37) n'imposent aucune contrainte sur les lignes de  $\boldsymbol{\Pi}$  qui correspondent à  $\mathbf{X}_1$ . En général, cependant, elles imposent des contraintes sur les lignes qui correspondent à  $\mathbf{X}_2$ .

Comme nous l'avons vu dans la section qui précédait, il y a équivalence entre la minimisation d'un déterminant tel que (18.34) sur lequel ne pèse aucune contrainte et l'usage des OLS. Dans ce cas, puisqu'aucune contrainte sur les lignes de  $\boldsymbol{\Pi}$  ne correspond à  $\mathbf{X}_1$ , nous pouvons employer les OLS pour estimer ces paramètres, et ensuite concentrer ce déterminant par rapport à ces paramètres. Ce faisant, le déterminant dans le membre de droite de (18.40) devient

$$|(\mathbf{Y}\boldsymbol{\Gamma} - \mathbf{XB})^\top \mathbf{M}_1 (\mathbf{Y}\boldsymbol{\Gamma} - \mathbf{XB})|,$$

où, comme d'habitude,  $\mathbf{M}_1$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(\mathbf{X}_1)$ .

Nous allons à présent introduire une notation nouvelle. Premièrement, notons  $\boldsymbol{\gamma}$  le vecteur  $[1 \vdots -\gamma_1]$ ; par conséquent,  $\mathbf{Y}\boldsymbol{\gamma} \equiv \mathbf{y} - \mathbf{Y}_1\gamma_1$ . Deuxièmement, notons  $\mathbf{Y}^* \mathbf{M}_1 \mathbf{Y}$ ,  $\mathbf{Y}_1^* \mathbf{M}_1 \mathbf{Y}_1$ , et  $\mathbf{X}^* \mathbf{M}_1 \mathbf{X}_2$ . On peut récrire le déterminant dans le membre de droite de (18.40) comme

$$\begin{vmatrix} (\mathbf{Y}^*\boldsymbol{\gamma})^\top(\mathbf{Y}^*\boldsymbol{\gamma}) & (\mathbf{Y}^*\boldsymbol{\gamma})^\top(\mathbf{Y}_1^* - \mathbf{X}^*\mathbf{B}_{22}) \\ (\mathbf{Y}_1^* - \mathbf{X}^*\mathbf{B}_{22})^\top(\mathbf{Y}^*\boldsymbol{\gamma}) & (\mathbf{Y}_1^* - \mathbf{X}^*\mathbf{B}_{22})^\top(\mathbf{Y}_1^* - \mathbf{X}^*\mathbf{B}_{22}) \end{vmatrix}. \quad (18.41)$$

Ce déterminant ne dépend que des paramètres  $\gamma$  et  $B_{22}$ . La prochaine étape consiste à concentrer par rapport aux paramètres de  $B_{22}$ , de manière à obtenir une expression qui ne dépend que de  $\gamma$ . Cela nécessitera un usage intensif du résultat suivant, qui est démontré dans l'Annexe A:

$$\begin{vmatrix} A^\top A & A^\top B \\ B^\top A & B^\top B \end{vmatrix} = |A^\top A| |B^\top M_A B|, \quad (18.42)$$

où, comme d'habitude,  $M_A \equiv I - A(A^\top A)^{-1}A^\top$ . Lorsque ce résultat est appliqué à (18.41), nous obtenons

$$(Y^* \gamma)^\top (Y^* \gamma) \left| (Y_1^* - X^* B_{22})^\top M_v (Y_1^* - X^* B_{22}) \right|, \quad (18.43)$$

où  $M_v$  désigne la matrice qui projette orthogonalement sur  $\mathcal{S}^\perp(v)$ , et  $v \equiv Y^* \gamma$ . Il n'existe qu'un seul déterminant dans (18.43), et non pas deux, parce que le premier est un scalaire.

Les paramètres  $B_{22}$  n'apparaissent que dans le second facteur de (18.43). Ce facteur est le déterminant de la matrice des sommes des carrés et des produits croisés des résidus du système des régressions entier

$$M_v Y_1^* = M_v X^* B_{22} + \text{résidus}.$$

Comme nous l'avons vu dans la section précédente, ce déterminant peut être minimisé en remplaçant  $B_{22}$  par son estimation, obtenue en appliquant les OLS à chaque équation séparément. La matrice des résidus ainsi produite est  $M_{M_v X^*} M_v Y_1^*$ , où  $M_{M_v X^*}$  désigne la projection sur le complément orthogonal de  $\mathcal{S}(M_v X^*)$ . Observons à présent que  $M_{M_v X^*} M_v = M_{v, X^*}$ , à savoir la matrice de projection associée au complément orthogonal de  $\mathcal{S}(v, X^*)$ . Conséquemment, le second facteur de (18.43), lorsqu'il est minimisé par rapport à  $B_{22}$ , est

$$\left| (Y_1^*)^\top M_{v, X^*} Y_1^* \right|. \quad (18.44)$$

On peut exploiter le fait que  $v$  et  $X^*$  apparaissent de manière symétrique dans (18.44) afin de faire dépendre (18.44) de  $\gamma$  uniquement à travers un scalaire. Considérons le déterminant

$$\begin{vmatrix} v^\top M_{X^*} v & v^\top M_{X^*} Y_1^* \\ (Y_1^*)^\top M_{X^*} v & (Y_1^*)^\top M_{X^*} Y_1^* \end{vmatrix}. \quad (18.45)$$

En utilisant (18.42), ce déterminant peut être factorisé tout comme (18.41). Nous aboutissons à

$$(v^\top M_{X^*} v) \left| (Y_1^*)^\top M_{v, X^*} Y_1^* \right|. \quad (18.46)$$

En faisant usage des définitions  $M_1 M_{X^*} = M_X$  et  $v = M_1 Y \gamma$ , (18.45) peut être récrit

$$\begin{vmatrix} \gamma^\top Y^\top M_X Y \gamma & \gamma^\top Y^\top M_X Y_1 \\ Y_1^\top M_X Y \gamma & Y_1^\top M_X Y_1 \end{vmatrix} = |\Gamma^\top Y^\top M_X Y \Gamma| = |Y^\top M_X Y|. \quad (18.47)$$

La première égalité est ici aisément vérifiée en exploitant l'expression (18.39) pour  $\mathbf{I}$  et les définitions de  $\boldsymbol{\gamma}$  et  $\mathbf{Y}$ ; souvenons-nous que  $\boldsymbol{\gamma}$  est la première colonne de  $\mathbf{I}$ . La seconde égalité est un résultat du fait que  $|\mathbf{I}| = 1$ . Elle implique que (18.47) ne dépend pas du tout de  $\mathbf{I}$ .

Enfin, nous pouvons maintenant écrire une expression simplifiée, qui, lorsqu'elle est minimisée par rapport à  $\boldsymbol{\gamma}$ , est égale à la valeur minimisée du déterminant originel (18.40). De (18.46) et (18.47), nous voyons que (18.44) est égal à

$$|(\mathbf{Y}_1^*)^\top \mathbf{M}_{v, X^*} \mathbf{Y}_1^*| = \frac{|\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|}{\mathbf{v}^\top \mathbf{M}_{X^*} \mathbf{v}} = \frac{|\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|}{\boldsymbol{\gamma}^\top \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \boldsymbol{\gamma}}.$$

Ainsi, en utilisant (18.43), le déterminant d'origine (18.40) doit être égal à

$$\frac{\mathbf{v}^\top \mathbf{v} |\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|}{\boldsymbol{\gamma}^\top \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \boldsymbol{\gamma}} = \frac{(\boldsymbol{\gamma}^\top \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} \boldsymbol{\gamma}) |\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|}{\boldsymbol{\gamma}^\top \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \boldsymbol{\gamma}} = \kappa |\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|, \quad (18.48)$$

où le scalaire  $\kappa$  a été défini implicitement comme

$$\kappa \equiv \frac{\boldsymbol{\gamma}^\top \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^\top \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \boldsymbol{\gamma}}. \quad (18.49)$$

Puisque  $|\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|$  ne dépend pas du tout de  $\boldsymbol{\gamma}$ , il y a équivalence entre la minimisation de (18.48) et la minimisation de  $\kappa$ . Ainsi, si nous pouvons minimiser (18.49) par rapport à  $\boldsymbol{\gamma}$ , nous pouvons obtenir des estimations LIML  $\hat{\boldsymbol{\gamma}}$  et une valeur associée de  $\kappa$ , disons  $\hat{\kappa}$ . Lorsque les estimations LIML sont obtenues de cette manière, on les appelle quelquefois estimations du **rapport de moindre variance**.

Avant de voir comment obtenir des estimations LIML  $\hat{\boldsymbol{\gamma}}$ , il nous faut dire quelques mots des conséquences de (18.48) et (18.49). En premier lieu, il devrait être évident que  $\hat{\kappa} \geq 1$ . Etant donné que  $\mathcal{S}(\mathbf{X}_1)$  est un sous-espace de  $\mathcal{S}(\mathbf{X})$ , le numérateur de (18.49) ne peut pas être inférieur au dénominateur pour tout  $\boldsymbol{\gamma}$  possible. En fait, pour une équation suridentifiée,  $\hat{\kappa}$  sera toujours supérieur à 1 avec des échantillons finis. En ce qui concerne une équation juste identifiée,  $\hat{\kappa}$  sera précisément égal à 1 parce que le nombre de paramètres à estimer est alors égal à  $k$ , le rang de  $\mathbf{X}$ . Ainsi, dans ce cas, il est possible de choisir  $\boldsymbol{\gamma}$  de sorte que le numérateur et le dénominateur de (18.49) soient égaux.

L'expression (18.48) implique que la valeur maximisée de la fonction de log-vraisemblance concentrée pour l'estimation LIML d'une unique équation est

$$-\frac{ng}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\kappa}) - \frac{n}{2} \log |\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|. \quad (18.50)$$

La valeur maximisée de la fonction de log-vraisemblance concentrée pour l'estimation ML de la forme réduite libre est

$$-\frac{ng}{2} \log(2\pi) - \frac{n}{2} \log |\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y}|.$$



Par conséquent une statistique LR portant sur les contraintes de suridentification implicites dans une seule équation structurelle est simplement  $n \log(\hat{\kappa})$ . Cette statistique de test fut proposée à l'origine par Anderson et Rubin (1950).

Il est aisé d'évaluer  $\hat{\kappa}$ . L'ensemble des conditions du premier ordre obtenu en dérivant (18.49) par rapport à  $\gamma$  est

$$2\mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} \gamma (\gamma^\top \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \gamma) - 2\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \gamma (\gamma^\top \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} \gamma) = \mathbf{0}.$$

Si nous divisons chaque membre de l'égalité par  $2\gamma^\top \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \gamma$ , nous aboutissons

$$\mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} \gamma - \kappa \mathbf{Y}^\top \mathbf{M}_X \mathbf{Y} \gamma = \mathbf{0}. \quad (18.51)$$

Un ensemble de conditions du premier ordre équivalent peut être établi en prémultipliant (18.51) par  $(\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{-1/2}$  et en insérant ce facteur multiplié par son inverse devant  $\gamma$ . Après manipulation, nous arrivons à

$$((\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{-1/2} \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} (\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{-1/2} - \kappa \mathbf{I})(\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{1/2} \gamma = \mathbf{0}.$$

Cet ensemble de conditions du premier ordre possède désormais la forme d'un problème classique de valeurs propres et vecteurs propres pour une matrice réelle symétrique (voir Annexe A). Il est clair désormais que  $\hat{\kappa}$  sera une valeur propre de la matrice

$$(\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{-1/2} \mathbf{Y}^\top \mathbf{M}_1 \mathbf{Y} (\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{-1/2} \quad (18.52)$$

et que  $(\mathbf{Y}^\top \mathbf{M}_X \mathbf{Y})^{1/2} \hat{\gamma}$  sera son vecteur propre associé. En réalité,  $\hat{\kappa}$  doit être la valeur propre *la plus petite*, du fait que c'est la plus faible valeur du rapport (18.49).

Alors, un moyen de calculer des estimations LIML consiste à trouver le vecteur propre (18.52) associé à la valeur propre la plus petite, et de là, à calculer  $\hat{\gamma}$ , qui sera  $[1 \vdots -\hat{\gamma}_1]$  si le premier élément est normalisé à 1. On peut ensuite obtenir  $\hat{\beta}_1$  en régressant  $\mathbf{y} - \mathbf{Y}_1 \hat{\gamma}_1$  sur  $\mathbf{X}_1$ . Une approche alternative se révèle pourtant plus simple et plus révélatrice. Considérons les conditions du premier ordre (18.51). Si nous les exprimons en termes de  $\mathbf{y}$  et  $\mathbf{Y}_1$  au lieu de  $\mathbf{Y}$ , et les évaluons avec les estimations LIML, nous pouvons les récrire sous la forme

$$\left( \begin{bmatrix} \mathbf{y}^\top \mathbf{M}_1 \mathbf{y} & \mathbf{y}^\top \mathbf{M}_1 \mathbf{Y}_1 \\ \mathbf{Y}_1^\top \mathbf{M}_1 \mathbf{y} & \mathbf{Y}_1^\top \mathbf{M}_1 \mathbf{Y}_1 \end{bmatrix} - \hat{\kappa} \begin{bmatrix} \mathbf{y}^\top \mathbf{M}_X \mathbf{y} & \mathbf{y}^\top \mathbf{M}_X \mathbf{Y}_1 \\ \mathbf{Y}_1^\top \mathbf{M}_X \mathbf{y} & \mathbf{Y}_1^\top \mathbf{M}_X \mathbf{Y}_1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ -\hat{\gamma}_1 \end{bmatrix} = \mathbf{0}.$$

Pour ce qui concerne les lignes correspondant à  $\mathbf{Y}_1$ , nous avons

$$\mathbf{Y}_1^\top (\mathbf{M}_1 - \hat{\kappa} \mathbf{M}_X) \mathbf{y} - \mathbf{Y}_1^\top (\mathbf{M}_1 - \hat{\kappa} \mathbf{M}_X) \mathbf{Y}_1 \hat{\gamma}_1 = \mathbf{0}.$$

En résolvant par rapport à  $\hat{\gamma}_1$ , nous obtenons

$$\hat{\gamma}_1 = (\mathbf{Y}_1^\top (\mathbf{M}_1 - \hat{\kappa} \mathbf{M}_X) \mathbf{Y}_1)^{-1} \mathbf{Y}_1^\top (\mathbf{M}_1 - \hat{\kappa} \mathbf{M}_X) \mathbf{y}.$$

Puisque  $\mathbf{X}_1 \in \mathcal{S}(\mathbf{X})$ ,  $\mathbf{M}_1 - \hat{\kappa}\mathbf{M}_X = \mathbf{M}_1(\mathbf{I} - \hat{\kappa}\mathbf{M}_X)$ . A l'aide de cette propriété et d'un peu d'algèbre, on peut montrer que  $\hat{\gamma}_1$  peut également se calculer suivant la formule (nous laissons la manipulation en qu'exercice)

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\gamma}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{Y}_1 \\ \mathbf{Y}_1^\top \mathbf{X}_1 & \mathbf{Y}_1^\top (\mathbf{I} - \hat{\kappa}\mathbf{M}_X) \mathbf{Y}_1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{y} \\ \mathbf{Y}_1^\top (\mathbf{I} - \hat{\kappa}\mathbf{M}_X) \mathbf{y} \end{bmatrix}, \quad (18.53)$$

qui fournit également  $\hat{\beta}_1$ . Alors si nous définissons  $\mathbf{Z}$  par  $[\mathbf{X}_1 \ \mathbf{Y}_1]$  et  $\boldsymbol{\delta}$  par  $[\beta_1 \ ; \ \gamma_1]$ , tout comme dans (18.18), (18.53) peut se récrire sous la forme très simple

$$\hat{\boldsymbol{\delta}} = (\mathbf{Z}^\top (\mathbf{I} - \hat{\kappa}\mathbf{M}_X) \mathbf{Z})^{-1} \mathbf{Z}^\top (\mathbf{I} - \hat{\kappa}\mathbf{M}_X) \mathbf{y}. \quad (18.54)$$

L'équation (18.53) est un moyen parmi d'autres d'écrire le LIML comme un membre des estimateurs de classe  $K$ ; voir Theil (1961) et Nagar (1959). L'équation (18.54) est un moyen encore plus simple d'arriver au même but. La classe  $K$  comprend tous les estimateurs que l'on peut écrire sous une de ces deux formes, mais avec un scalaire  $K$  arbitraire à la place de  $\hat{\kappa}$ . Nous employons la notation  $K$  plutôt que la notation plus conventionnelle  $k$  pour désigner ce scalaire afin d'éviter la confusion avec le nombre de variables exogènes dans le système. L'estimateur LIML est ainsi un estimateur de la classe  $K$ , avec la paramétrisation  $K = \hat{\kappa}$ . Identiquement, comme (18.54) le montre clairement, l'estimateur 2SLS est un estimateur de la classe  $K$  avec la paramétrisation  $K = 1$ , et celui des OLS est également un estimateur de la classe  $K$  avec la paramétrisation  $K = 0$ . Puisque pour une équation structurelle juste identifiée,  $\hat{\kappa} = 1$ , il découle immédiatement de (18.54) que les estimateurs LIML et 2SLS se confondent dans ce cas particulier.

On peut montrer que les estimateurs de la classe  $K$  sont convergents lorsque  $K$  tend vers 1 asymptotiquement à un taux plus fort que  $n^{-1/2}$ ; voir Schmidt (1976), parmi d'autres auteurs. Bien que la convergence du LIML provienne de résultats généraux sur les estimateurs ML, il reste intéressant de voir comment ce résultat pour la classe  $K$  s'y applique. Nous avons déjà vu que  $n \log(\hat{\kappa})$  est la statistique de test LR pour l'hypothèse nulle de pertinence des contraintes de suridentification sur l'équation structurelle. Un développement de Taylor sur le logarithme nous montre que  $n \log(\hat{\kappa}) \cong n(\hat{\kappa} - 1)$ . Puisque cette statistique de test suit asymptotiquement une loi du  $\chi^2$ , elle doit être  $O(1)$ , de sorte que  $\hat{\kappa} - 1$  doit être  $O(n^{-1})$ . Ceci établit la convergence du LIML.

Il existe de nombreux autres estimateurs de la classe  $K$ . Par exemple, Sawa (1973) suggéra un moyen de modifier l'estimateur 2SLS pour réduire son biais, et Fuller (1977) et Morimune (1978, 1983) suggérèrent des versions modifiées de l'estimateur LIML. L'estimateur de Fuller, qui est le plus simple d'entre eux, utilise la paramétrisation  $K = \hat{\kappa} - \alpha/(n - k)$ , où  $\alpha$  est une constante positive que choisit l'expérimentateur. Un choix judicieux est  $\alpha = 1$ , puisqu'il produit des estimations approximativement non biaisées. Par

contraste avec l'estimateur LIML qui ne possède aucun moment fini (voir Mariano (1982) et Phillips (1983) sur ce point), tous les moments de l'estimateur modifié de Fuller sont finis à condition que l'échantillon soit suffisamment important.

Il est possible d'estimer la matrice de covariance du vecteur  $\hat{\delta}$  des estimations de la classe  $K$  de différentes façons. La plus naturelle consiste à utiliser

$$\hat{\sigma}^2(\mathbf{Z}^\top(\mathbf{I} - \hat{\kappa}\mathbf{M}_X)\mathbf{Z})^{-1}, \quad (18.55)$$

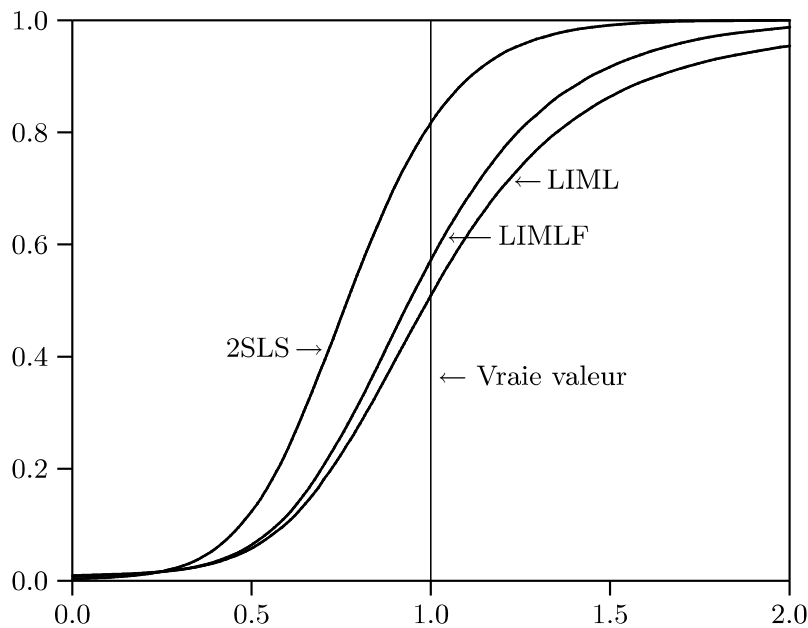
où

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{Z}\hat{\delta})^\top(\mathbf{y} - \mathbf{Z}\hat{\delta}).$$

Les statistiques de test de Wald pour les contraintes sur  $\gamma_1$  et  $\beta_1$ , et parmi elles les  $t$  de Student asymptotiques, peuvent se calculer à l'aide de (18.55) de la manière habituelle. Toutefois, il est sans doute préférable d'employer des statistiques LR, étant donné leur invariance à la reparamétrisation, mais aussi compte tenu de leur facilité de calcul à partir de la fonction de log-vraisemblance concentrée (18.50).

Le résultat selon lequel les estimateurs de la classe  $K$  sont convergents lorsque  $K$  tend asymptotiquement vers 1 à un taux approprié peut suggérer que les 2SLS possèdent de meilleures propriétés avec des échantillons finis que le LIML. Après tout, pour les 2SLS,  $K$  est identiquement égal à 1, alors que pour le LIML,  $K = \hat{\kappa}$ , et  $\hat{\kappa}$  est toujours supérieur à 1 avec des échantillons finis. Le résultat selon lequel le LIML ne possède pas de moment fini peut également suggérer que cet estimateur est plus pauvre que celui des 2SLS, puisque, comme nous l'avons vu dans la Section 7.5, l'estimateur des 2SLS possèdent autant de moments finis qu'il y a de contraintes de suridentification. D'un autre côté, il apparaît que dans de nombreux cas, les 2SLS possèdent en fait de piètres qualités face au LIML à de multiples égards. Anderson, Kunitomo, et Sawa (1982), par exemple, exposent des résultats analytiques qui montrent que le LIML converge vers sa distribution asymptotique normale beaucoup plus rapidement que ne le font les 2SLS. Contrairement à la distribution de l'estimateur 2SLS, dont nous avons vu qu'elle est sévèrement biaisée dans certains cas, la distribution de l'estimateur LIML est généralement centrée sur une valeur proche de la véritable valeur. Mais, étant donné que cette dernière distribution ne possède pas de moment fini, nous ne pouvons pas conclure au moindre biais de l'estimateur LIML.

La Figure 18.1 donne une illustration du fonctionnement du LIML avec des échantillons finis. Elle montre les distributions de l'estimateur 2SLS, l'estimateur LIML, et l'estimateur modifié de Fuller avec  $\alpha = 1$  (noté LIMLF sur la figure) dans le cas examiné précédemment dans la Section 7.5. La présence de 6 contraintes de suridentification et de seulement 25 observation explique la divergence importante pour chaque estimateur par rapport à sa distribution asymptotique. Dans ce cas, l'estimateur 2SLS est sévèrement biaisé vers le bas. Par ailleurs, l'estimateur LIML semble être pratiquement



**Figure 18.1** Distributions des estimateurs 2SLS et LIML

sans biais dans le sens où sa médiane est très proche de la véritable valeur de 1. La distribution de l'estimateur modifié de Fuller se situe généralement entre celles des estimateurs 2SLS et LIML. Sa queue de distribution supérieure est beaucoup plus fine que celle du LIML, mais sa médiane est quelque peu inférieure à la véritable valeur.

Dans la pratique, il n'est pas toujours aisé de décider quel estimateur de la classe  $K$  utiliser. Mariano (1982) aborde un certain nombre de résultats analytiques et donne des conseils sur l'opportunité d'une performance meilleure du LIML par rapport aux 2SLS. Il faudrait éviter d'employer ce dernier lorsque le nombre des contraintes de suridentification est important, par exemple. Cependant, cela dépend énormément des caractéristiques intrinsèques du modèle et des données que l'on utilise. Si les résultats des 2SLS et du LIML sont très proches, alors le choix entre les deux est peu important. S'ils sont relativement différents, toutefois, ce choix devient important. Sans doute la meilleure chose à faire dans ces circonstances consiste à réaliser des expériences Monte Carlo, qui sont typiquement conçues pour départager les performances relatives des différents estimateurs pour le modèle et les données en cause; se reporter au Chapitre 21.

## 18.6 LES TRIPLES MOINDRES CARRÉS

La dernière des quatre méthodes principales pour l'estimation des modèles d'équations simultanées dont nous allons discuter est celle des **triples moindres carrés**, ou **3SLS**. Tout comme le FIML, la méthode des 3SLS est une

méthode systématique, pour laquelle tous les paramètres du modèle sont estimés conjointement. Ainsi que son nom le suggère, on peut calculer les 3SLS en trois étapes. Les deux premières sont celles des 2SLS classiques, appliquées à chaque équation du système séparément. La troisième étape est alors essentiellement la même que l'étape terminale de l'estimation par GLS faisables d'un système SUR (Section 9.7). La méthode fut proposée par Zellner et Theil (1962).

Le moyen le plus simple de dériver l'estimateur des 3SLS, ainsi que ses propriétés asymptotiques, consiste à appliquer les principes de la méthode des moments généralisée au système des modèles d'équations simultanées linéaires (18.01). Pour l'observation  $t$ , ce système peut se mettre sous la forme

$$\mathbf{Y}_t \boldsymbol{\Gamma} = \mathbf{X}_t \mathbf{B} + \mathbf{U}_t.$$

L'hypothèse selon laquelle toutes les variables dans  $\mathbf{X}$  sont soit exogènes soit prédéterminées implique que, pour toutes les observations  $t$ ,

$$E(\mathbf{Y}_t \boldsymbol{\Gamma} - \mathbf{X}_t \mathbf{B} | \mathbf{X}_t) = \mathbf{0}.$$

On interprète immédiatement les égalités comme des conditions portant sur les moments conditionnels au sens du Chapitre 17. Puisque, comme nous l'avons vu dans la Section 18.3, les variables exogènes constituent des instruments efficaces pour les 2SLS si les aléas sont homoscédastiques et indépendants en série, il semble raisonnable d'envisager l'ensemble suivant de conditions du premier ordre:

$$E(\mathbf{X}_t^\top (\mathbf{Y}_t \boldsymbol{\Gamma} - \mathbf{X}_t \mathbf{B})) = \mathbf{0}. \quad (18.56)$$

Etant donné que  $\mathbf{X}_t$  possède  $k$  composantes et  $\mathbf{Y}_t \boldsymbol{\Gamma} - \mathbf{X}_t \mathbf{B}$  en possède  $g$ , il y a en tout  $gk$  conditions portant sur les moments. Si la condition d'ordre pour l'identification est satisfaite avec une égalité, il y aurait exactement  $gk$  paramètres à estimer. Ainsi (18.56) fournit toujours au moins autant de conditions portant sur les moments qu'il y a de paramètres dans le système, et même davantage si le système est suridentifié. Bien évidemment, l'utilité réelle de ces conditions sur les moments dans le processus d'identification des paramètres dépend asymptotiquement de la validité de la condition de rang.

Il est pratique d'ordonner différemment les éléments de la matrice de dimension  $k \times g$  (18.56) pour en faire un vecteur de dimension  $gk$ . En premier lieu, exprimons chaque équation du système dans une notation comparable à celle de (18.18):

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\delta}_i + \mathbf{u}_i, \quad \text{pour } i = 1, \dots, g,$$

où la matrice de régresseurs  $\mathbf{Z}_i$  qui apparaît dans l'équation  $i$  est  $[\mathbf{X}_i \ \mathbf{Y}_i]$ , avec  $k_i$  variables exogènes  $\mathbf{X}_i$  incluses et  $g_i$  variables endogènes  $\mathbf{Y}_i$  incluses, et où le vecteur de paramètres de dimension  $(k_i + g_i)$   $\boldsymbol{\delta}_i$  est  $[\boldsymbol{\beta}_i \ ; \ \boldsymbol{\gamma}_i]$ . Définissons alors le vecteur ligne  $\mathbf{F}_t$  composé de  $gk$  éléments comme:

$$\mathbf{F}_t \equiv [u_{t1} \mathbf{X}_t \ \cdots \ u_{tg} \mathbf{X}_t],$$

où  $u_{ti} \equiv y_{ti} - (\mathbf{Z}_i)_t \boldsymbol{\delta}_i$ . Chaque composante de  $\mathbf{F}_t$  est la contribution de l'observation  $t$  à un des moments empiriques provenant de (18.56). La matrice  $\mathbf{F}$  de dimension  $n \times gk$  est définie pour avoir une ligne type  $\mathbf{F}_t$ .

Pour obtenir des estimations GMM, il est nécessaire de trouver une estimation de la matrice de covariance des  $gk$  moments (18.56). Nous ferons les mêmes hypothèses préliminaires sur les aléas que pour le FIML et le LIML. Nous supposons que chaque vecteur  $\mathbf{u}_i$  est homoscédastique et indépendant en série (l'hypothèse d'homoscédasticité sera relâchée plus tard). Nous supposons également que, pour chaque observation  $t$ , les  $u_{ti}$  sont corrélés entre eux, avec une matrice de covariance contemporaine de dimension  $g \times g$   $\boldsymbol{\Sigma}$ , indépendante de  $t$ . Nous noterons  $\sigma_{ij}$  un élément type de  $\boldsymbol{\Sigma}$  et  $\sigma^{ij}$  un élément type de  $\boldsymbol{\Sigma}^{-1}$ .

Il est relativement aisé de trouver la matrice de covariance du vecteur des moments empiriques  $\mathbf{F}^\top \boldsymbol{\iota}$ . C'est

$$\begin{aligned} E(\mathbf{F}^\top \boldsymbol{\iota} \mathbf{F}) &= \sum_{t=1}^n E(\mathbf{F}_t^\top \mathbf{F}_t) \\ &= \sum_{t=1}^n E[u_{t1} \mathbf{X}_t \cdots u_{tg} \mathbf{X}_t]^\top [u_{t1} \mathbf{X}_t \cdots u_{tg} \mathbf{X}_t]. \end{aligned} \quad (18.57)$$

La dernière expression dans (18.57) est une matrice de dimension  $gk \times gk$  qui apparaît sous une forme plus lisible lorsqu'elle est partitionnée, chaque bloc étant de dimension  $k \times k$ . Pour chaque  $t$ ,  $E(u_{ti} u_{tj}) = \sigma_{ij}$ . Parce que les éléments de  $\sigma_{ij}$  ne dépendent pas de  $t$ , nous obtenons

$$\begin{bmatrix} \sigma_{11} \mathbf{X}^\top \mathbf{X} & \cdots & \sigma_{1g} \mathbf{X}^\top \mathbf{X} \\ \vdots & \ddots & \vdots \\ \sigma_{g1} \mathbf{X}^\top \mathbf{X} & \cdots & \sigma_{gg} \mathbf{X}^\top \mathbf{X} \end{bmatrix}, \quad (18.58)$$

c'est-à-dire une matrice dont le bloc type est  $\sigma_{ij} \mathbf{X}^\top \mathbf{X}$ . Afin de construire une fonction critère comparable à (17.54) et avec laquelle nous pourrions obtenir des estimations des paramètres vectoriels  $\boldsymbol{\delta}_i$ ,  $i = 1, \dots, g$ , nous aurons besoin d'inverser la matrice (18.58). La structure en bloc de (18.58) facilite cette manipulation. On peut vérifier facilement par une simple multiplication de matrices partitionnées que l'inverse est une matrice dont le bloc type est  $\sigma^{ij} (\mathbf{X}^\top \mathbf{X})^{-1}$  (souvenons-nous que  $\sigma^{ij}$  est un élément type de  $\boldsymbol{\Sigma}^{-1}$ ).

Il est pratique d'exprimer le vecteur des moments empiriques  $\mathbf{F}^\top \boldsymbol{\iota}$  sous une forme partitionnée comparable à (18.58), comme une fonction des données et des paramètres du modèle. Le résultat est un vecteur avec l'élément type  $\mathbf{X}^\top (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}_i)$ , pour  $i = 1, \dots, g$ :

$$\mathbf{F}^\top \boldsymbol{\iota} = \begin{bmatrix} \mathbf{X}^\top (\mathbf{y}_1 - \mathbf{Z}_1 \boldsymbol{\delta}_1) \\ \vdots \\ \mathbf{X}^\top (\mathbf{y}_g - \mathbf{Z}_g \boldsymbol{\delta}_g) \end{bmatrix}. \quad (18.59)$$

Alors, si nous élaborons une forme quadratique à partir du vecteur (18.59) et de la matrice (18.58), nous aboutissons à la fonction critère

$$\begin{aligned} & \sum_{i=1}^g \sum_{j=1}^g \sigma^{ij} (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}_i)^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\delta}_j) \\ &= \sum_{i=1}^g \sum_{j=1}^g \sigma^{ij} (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}_i)^\top \mathbf{P}_X (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\delta}_j). \end{aligned} \quad (18.60)$$

Puisque nous supposons tacitement qu'il n'existe aucune contrainte d'équations croisées, les paramètres  $\boldsymbol{\delta}_i$  n'apparaissent que dans le résidu de l'équation  $i$ . Ainsi les conditions du premier ordre pour un minimum de (18.60) peuvent s'écrire assez simplement comme

$$\sum_{j=1}^g \sigma^{ij} \mathbf{Z}_i^\top \mathbf{P}_X (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\delta}_j) = \mathbf{0}, \quad \text{pour } i = 1, \dots, g. \quad (18.61)$$

Afin de rendre (18.61) opérationnelle, nous avons besoin d'estimer la matrice de covariance des aléas,  $\boldsymbol{\Sigma}$ . Dans le cas du modèle SUR, nous pourrions employer les OLS pour chaque équation individuellement. Puisque les OLS sont non convergents pour les modèles d'équations simultanées, nous employons à la place les 2SLS sur chaque équation. Ainsi les deux premières "étapes" des 3SLS correspondent exactement aux deux étapes des 2SLS, appliqué à chaque équation de (18.01). Les covariances des aléas sont alors estimés à partir des résidus 2SLS:

$$\tilde{\sigma}_{ij} = \frac{1}{n} \sum_{t=1}^n \tilde{u}_{ti} \tilde{u}_{tj}. \quad (18.62)$$

Bien sûr, ces résidus doivent correspondre aux véritables résidus 2SLS, et non aux résidus de l'estimation OLS de seconde étape: voir la Section 7.5. Nous voyons donc que les estimateurs 3SLS,  $\tilde{\boldsymbol{\delta}}_1$  à  $\tilde{\boldsymbol{\delta}}_g$  doivent conjointement résoudre les conditions du premier ordre:

$$\sum_{j=1}^g \tilde{\sigma}^{ij} \mathbf{Z}_i^\top \mathbf{P}_X (\mathbf{y}_j - \mathbf{Z}_j \tilde{\boldsymbol{\delta}}_j) = \mathbf{0}. \quad (18.63)$$

La solution est aisée à formuler. Si  $\boldsymbol{\delta} \equiv [\boldsymbol{\delta}_1 : \dots : \boldsymbol{\delta}_g]$  et si les matrices entre crochets désignent les matrices partitionnées caractérisées par l'élément type à l'intérieur du crochet, l'estimateur 3SLS  $\tilde{\boldsymbol{\delta}}$  se met sous la forme compacte

$$\tilde{\boldsymbol{\delta}} = [\tilde{\sigma}^{ij} \mathbf{Z}_i^\top \mathbf{P}_X \mathbf{Z}_j]^{-1} \left[ \sum_{j=1}^g \tilde{\sigma}^{ij} \mathbf{Z}_i^\top \mathbf{P}_X \mathbf{y}_j \right]. \quad (18.64)$$

L'écriture de l'estimateur 3SLS dans une notation qui utilise les produits de Kronecker est plus fréquente; consulter la plupart des ouvrages d'économétrie. Bien que les produits de Kronecker soient bien souvent très utiles (Magnus et Neudecker, (1988)), nous préférons la notation compacte de (18.64).

L'estimateur 3SLS est intimement relié à la fois à celui des 2SLS et à celui des GLS pour les modèles SUR multivariés pour lequel les variables explicatives sont toutes exogènes ou prédéterminées. Si nous supposons que  $\Sigma$  est proportionnelle à une matrice identité, les conditions (18.63) se ramènent à

$$\tilde{\sigma}^{ii} \mathbf{Z}_i^\top \mathbf{P}_X (\mathbf{y}_i - \mathbf{Z}_i \boldsymbol{\delta}_i) = \mathbf{0},$$

et ces conditions sont équivalentes aux conditions équation par équation des 2SLS. Ainsi les 3SLS et les 2SLS seront asymptotiquement (mais pas numériquement) équivalents lorsque les aléas contemporains de la forme structurelle sont non corrélés. Il est également aisé de voir que l'estimateur SUR pour les modèles linéaires est juste un cas particulier de l'estimateur 3SLS. Etant donné que tous les régresseurs peuvent servir en tant qu'instruments dans le cas SUR, il n'est plus du tout besoin d'employer les 2SLS en première étape. En correspondance, le fait que chaque matrice de régresseur  $\mathbf{Z}_i$  soit une sous-matrice de la matrice de tous les régresseurs,  $\mathbf{X}$ , implique que  $\mathbf{P}_X \mathbf{Z}_i = \mathbf{Z}_i$ . Ainsi (18.63) se ramène à

$$\sum_{j=1}^g \tilde{\sigma}^{ij} \mathbf{Z}_i^\top (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\delta}_j) = \mathbf{0},$$

et c'est précisément ce que deviennent les équations définissantes (9.54) dans le cas linéaire pour l'estimateur des GLS faisables d'un système SUR sans contrainte d'équations croisées. Nous voyons que la relation entre 3SLS et les 2SLS équation par équation est identique à celle qu'il existe entre l'estimation SUR par GLS faisables et l'estimation OLS équation par équation.

Sur la base de (18.64), il est naturel de penser que l'estimation de la matrice de covariance de l'estimateur 3SLS peut être estimée par

$$[\tilde{\sigma}^{ij} \mathbf{Z}_i^\top \mathbf{P}_X \mathbf{Z}_j]^{-1}. \quad (18.65)$$

C'est en réalité le cas, comme on peut le montrer assez facilement à l'aide du résultat général (17.55) pour l'estimation GMM. Nous avons vu que pour  $\tilde{\Phi}^{-1}$  dans cette expression nous devons employer la matrice dont l'élément type est  $\tilde{\sigma}^{ij} (\mathbf{X}^\top \mathbf{X})^{-1}$ . Pour  $\tilde{\mathbf{D}}$ , la matrice des dérivées des moments empiriques par rapport aux paramètres du modèle, nous voyons que la matrice adéquate doit être bloc diagonale, avec des blocs types définis par  $-\mathbf{X}^\top \mathbf{Z}_i$ . (Nous ne considérons pas volontairement les facteurs des puissances de  $n$ .) Puisque nous traitons d'un système linéaire,  $\tilde{\mathbf{D}}$  ne dépend d'aucun paramètre estimé. Ainsi une estimation appropriée de la matrice de covariance asymptotique est donnée par l'inverse de la matrice dont le bloc type est

$$\mathbf{Z}_i^\top \mathbf{X} \tilde{\sigma}^{ij} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}_j = \tilde{\sigma}^{ij} \mathbf{Z}_i^\top \mathbf{P}_X \mathbf{Z}_j,$$



ce qui correspond précisément à (18.65).

Puisque le modèle d'équations simultanées (18.01) est équivalent à la forme réduite contrainte (18.02), on peut raisonnablement se demander pourquoi un estimateur tel que celui des 3SLS ne peut pas être obtenu simplement à partir de (18.02), étant donné que sa forme est précisément celle d'un système SUR. La réponse est, bien sûr, que cela est possible. Cependant, à moins que chaque équation ne soit juste identifiée, les contraintes seront non linéaires. Cette approche a été essentiellement utilisée par Chamberlain (1984). L'avantage de l'approche que nous suivons est qu'elle évite les difficultés associées au traitement des contraintes non linéaires.

Une autre similitude entre les estimations 3SLS et SUR est que les deux sont numériquement équivalentes à la procédure équation par équation si chaque équation est juste identifiée. Pour les systèmes SUR, cela signifie simplement que tous les régresseurs se confondent avec des variables explicatives dans chaque équation (sinon, il existerait des contraintes de suridentification impliquées par la nécessaire orthogonalité entre les aléas des équations où certains régresseurs sont absents et les régresseurs absents et inclus dans l'équation). Nous avons vu dans la Section 9.8, à travers le Théorème de Kruskal, que les estimations SUR sont numériquement identiques aux estimations OLS équation par équation dans ce cas. C'est un bon exercice que de montrer la validité du même résultat dans le contexte 3SLS.

Si nous supposons que les aléas contenus dans la matrice  $\mathbf{U}$  de (18.01) sont normalement distribués, les propriétés asymptotiques de toutes les procédures d'estimation ML garantissent l'efficacité asymptotique de l'estimateur FIML. Il est par conséquent naturel de se demander si l'estimateur 3SLS partage la propriété asymptotique d'efficacité avec le FIML, et la réponse est, comme nous le verrons assez directement, affirmative. Nous pourrions directement obtenir une démonstration de ce résultat si nous avions une expression de la matrice de covariance asymptotique de l'estimateur FIML, que nous pourrions comparer à (18.65). Toutefois, nous préférons ne pas obtenir une telle expression dans la Section 18.4, parce qu'un moyen très simple d'obtenir une estimation de la matrice de covariance FIML consiste à utiliser l'estimation 3SLS (18.65), évaluée avec les estimations FIML. Au lieu de cela, notre démonstration de l'équivalence asymptotique entre les 3SLS et le FIML se base sur le fait que l'estimateur FIML peut s'interpréter comme un estimateur des variables instrumentales.

Ce résultat, que Hausman (1975) démontra le premier, est d'un intérêt considérable en lui-même, du fait qu'il fournit des instruments optimaux associés à l'estimation ML du système (18.01). Comme nous pouvions nous y attendre, on peut les trouver en considérant les conditions du premier ordre pour la maximisation de la fonction de log-vraisemblance, que nous envisageons sous la forme (18.28). Si nous notons  $\mathbf{F}_i$  ou  $\mathbf{B}_i$  la colonne  $i$  de  $\mathbf{F}$  ou  $\mathbf{B}$ , respectivement, et notons une fois de plus  $\sigma^{ij}$  l'élément type de  $\mathbf{\Sigma}^{-1}$ , alors

(18.28) peut s'exprimer comme

$$\begin{aligned} \ell(\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Sigma}) = & -\frac{ng}{2} \log(2\pi) + n \log |\det \mathbf{\Gamma}| - \frac{n}{2} \log |\mathbf{\Sigma}| \\ & - \frac{1}{2} \sum_{t=1}^n \sum_{i=1}^g \sum_{j=1}^g \sigma^{ij} (\mathbf{Y}_t \mathbf{\Gamma}_i - \mathbf{X}_t \mathbf{B}_i) (\mathbf{Y}_t \mathbf{\Gamma}_j - \mathbf{X}_t \mathbf{B}_j). \end{aligned} \quad (18.66)$$

La difficulté majeure dans l'explicitation des conditions du premier ordre pour un maximum de (18.66) est que  $\mathbf{B}$  et  $\mathbf{\Gamma}$  sont contraintes à posséder de nombreux éléments nuls de sorte qu'un seul élément de  $\mathbf{\Gamma}$  est égal à 1. Par conséquent, nous ne pourrions annuler les dérivées de (18.66) par rapport à aux éléments de  $\mathbf{\Gamma}$  et  $\mathbf{B}$  qui sont ainsi contraints. Pour contourner la difficulté, nous pouvons tout d'abord développer une matrice des dérivées partielles de  $\ell(\mathbf{B}, \mathbf{\Gamma}, \mathbf{\Sigma})$  par rapport à  $\mathbf{B}$  qui aura exactement la même forme que la matrice  $\mathbf{B}$ . Nous signifions que l'élément  $ij$  de la matrice des dérivées partielles sera égal à la dérivée partielle de  $\ell$  par rapport à l'élément  $ij$  de la matrice  $\mathbf{B}$ . Nous pouvons exécuter une opération similaire pour  $\mathbf{\Gamma}$  et annuler uniquement les éléments pertinents des deux matrices de dérivées.

La matrice  $\mathbf{B}$  n'apparaît que dans le dernier terme de (18.66), aussi pouvons-nous nous focaliser uniquement sur ce terme pour l'instant. Il est commode de calculer la matrice des dérivées partielles élément par élément et d'ordonner ces dérivées par la suite dans une matrice de dimension  $k \times g$ . Puisque chaque facteur dans le dernier terme de (18.66) est un scalaire, chaque dérivée est aisément calculable. Par rapport à l'élément  $ij$ , nous obtenons

$$\sum_{t=1}^n \sum_{m=1}^g \sigma^{im} \mathbf{X}_{tj} (\mathbf{Y}_t \mathbf{\Gamma}_m - \mathbf{X}_t \mathbf{B}_m). \quad (18.67)$$

Nous souhaitons trouver une matrice dont l'élément  $ij$  est (18.67). Puisque  $j$  est l'indice associé à l'élément  $\mathbf{X}_{tj}$ , nous pouvons développer la colonne  $j$  de ladite matrice en ordonnant les éléments  $\mathbf{X}_{tj}$  en colonne. Cela donne

$$\begin{aligned} & \sum_{t=1}^n \sum_{m=1}^g \sigma^{im} \mathbf{X}_t^\top (\mathbf{Y}_t \mathbf{\Gamma}_m - \mathbf{X}_t \mathbf{B}_m) \\ &= \sum_{m=1}^g \sigma^{im} \mathbf{X}^\top (\mathbf{Y} \mathbf{\Gamma}_m - \mathbf{X} \mathbf{B}_m) \\ &= \mathbf{X}^\top (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B}) (\mathbf{\Sigma}^{-1})_i, \end{aligned} \quad (18.68)$$

où  $(\mathbf{\Sigma}^{-1})_i$  est la  $i^{\text{ième}}$  colonne de  $\mathbf{\Sigma}^{-1}$ . Observons maintenant que les expressions successives dans (18.68) sont des vecteurs de dimension  $k$ . Pour conclure cette manipulation, il nous faut concaténer ces vecteurs pour former une matrice de dimension  $k \times g$ , et il est désormais évident que cette matrice est  $\mathbf{X}^\top (\mathbf{Y} \mathbf{\Gamma} - \mathbf{X} \mathbf{B}) \mathbf{\Sigma}^{-1}$ .

Il nous faut maintenant calculer les dérivées (18.66) par rapport à la matrice de dimension  $g \times g$   $\mathbf{\Gamma}$ . Des opérations identiques à celles menées pour  $\mathbf{B}$  montrent que la matrice des dérivées par rapport au dernier terme de (18.66) est

$$-\mathbf{Y}^\top(\mathbf{Y}\mathbf{\Gamma} - \mathbf{X}\mathbf{B})\mathbf{\Sigma}^{-1}.$$

Cette matrice est de dimension  $g \times g$ , ce qui est cohérent. Mais  $\mathbf{\Gamma}$  apparaît également à travers son déterminant dans le second terme de (18.66). Souvenons-nous (ou bien consultons l'Annexe A) que la dérive du logarithme du déterminant d'une matrice par rapport à l'élément  $ij$  de cette matrice est l'élément  $ji$  de l'inverse de la matrice. Par conséquent, la matrice des dérivées partielles correspondant à  $\mathbf{\Gamma}$  est

$$n(\mathbf{\Gamma}^{-1})^\top - \mathbf{Y}^\top(\mathbf{Y}\mathbf{\Gamma} - \mathbf{X}\mathbf{B})\mathbf{\Sigma}^{-1}. \quad (18.69)$$

Nous pouvons aboutir à une expression plus pratique que (18.69) en utilisant les conditions du premier ordre pour les éléments de la matrice de covariance  $\mathbf{\Sigma}$ . De (18.29), nous voyons que ces conditions donnent

$$\hat{\mathbf{\Sigma}} = n^{-1}(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}}), \quad (18.70)$$

où  $\hat{\mathbf{\Sigma}}$ ,  $\hat{\mathbf{\Gamma}}$ , et  $\hat{\mathbf{B}}$  désignent des estimations FIML. Si nous prémultiplions cette équation par  $n\hat{\mathbf{\Sigma}}^{-1}$ , la postmultiplions par  $\hat{\mathbf{\Gamma}}^{-1}$ , et la transposons, nous arrivons à

$$n(\hat{\mathbf{\Gamma}}^{-1})^\top = \mathbf{Y}^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})\hat{\mathbf{\Sigma}}^{-1} - (\hat{\mathbf{\Gamma}}^{-1})^\top \hat{\mathbf{B}}^\top \mathbf{X}^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})\hat{\mathbf{\Sigma}}^{-1}. \quad (18.71)$$

Puisque  $\mathbf{X}\hat{\mathbf{B}}\hat{\mathbf{\Gamma}}^{-1}$  est la matrice des valeurs ajustées de l'estimation de la forme réduite contrainte, nous la noterons  $\hat{\mathbf{Y}}$ : cela simplifiera la notation et aura le mérite de clarifier l'analyse ultérieure. Ainsi (18.71) peut s'écrire

$$n(\hat{\mathbf{\Gamma}}^{-1})^\top = \mathbf{Y}^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})\hat{\mathbf{\Sigma}}^{-1} - \hat{\mathbf{Y}}^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})\hat{\mathbf{\Sigma}}^{-1}.$$

Par suite, la matrice (18.69), évaluée avec les estimations ML, devient

$$-\hat{\mathbf{Y}}^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})\hat{\mathbf{\Sigma}}^{-1}.$$

Nous pouvons, après tant d'efforts, sélectionner les éléments de deux matrices de dérivées partielles qui sont véritablement nuls lorsque nous les évaluons avec les estimations ML. Les paramètres qui apparaissent dans l'équation  $i$  proviennent de la colonne  $i$  des matrices  $\mathbf{\Gamma}$  et  $\mathbf{B}$ , et les dérivées partielles correspondantes proviennent des colonnes  $i$  des matrices de dérivées partielles. En ce qui concerne la matrice  $\mathbf{B}$ , cette colonne est  $\mathbf{X}^\top(\mathbf{Y}\hat{\mathbf{\Gamma}} - \mathbf{X}\hat{\mathbf{B}})(\hat{\mathbf{\Sigma}}^{-1})_i$ . Nous souhaitons sélectionner dans cette colonne uniquement les lignes pour lesquelles l'élément correspondant de  $\mathbf{B}_i$  est non contraint,

c'est-à-dire les éléments correspondant à la matrice de dimension  $n \times k_i$   $\mathbf{X}_i$ . Puisque pour sélectionner les lignes d'un produit matriciel, il nous suffit de sélectionner les lignes correspondant au facteur le plus à gauche, les éléments nuls sont ceux du vecteur de dimension  $k_i$   $\mathbf{X}_i^\top(\mathbf{Y}\hat{\Gamma} - \mathbf{X}\hat{\mathbf{B}})(\hat{\Sigma}^{-1})_i$ .

Par un raisonnement en tous points identique, nous trouvons que, pour chaque  $i = 1, \dots, g$ , le vecteur  $\hat{\mathbf{Y}}_i^\top(\mathbf{Y}\hat{\Gamma} - \mathbf{X}\hat{\mathbf{B}})(\hat{\Sigma}^{-1})_i$  de dimension  $g_i$  est nul, où  $\hat{\mathbf{Y}}_i$  ne contient que les colonnes de  $\hat{\mathbf{Y}}$  qui correspondent à la matrice  $\mathbf{Y}_i$  des variables endogènes incluses en tant que régresseurs dans l'équation  $i$ . Si nous définissons  $\hat{\mathbf{Z}}_i \equiv [\mathbf{X}_i \quad \hat{\mathbf{Y}}_i]$ , alors nous pouvons écrire toutes les conditions du premier ordre correspondant aux paramètres de la  $i^{\text{ième}}$  équation sous la forme

$$\hat{\mathbf{Z}}_i^\top(\mathbf{Y}\hat{\Gamma} - \mathbf{X}\hat{\mathbf{B}})(\Sigma^{-1})_i = \mathbf{0}.$$

Ces conditions peuvent se simplifier grandement. Remarquons que

$$\begin{aligned} (\mathbf{Y}\hat{\Gamma} - \mathbf{X}\hat{\mathbf{B}})(\hat{\Sigma}^{-1})_i &= \sum_{j=1}^g \hat{\sigma}^{ij}(\mathbf{Y}\hat{\Gamma}_j - \mathbf{X}\hat{\mathbf{B}}_j) \\ &= \sum_{j=1}^g \hat{\sigma}^{ij}(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j). \end{aligned}$$

L'ensemble complet des conditions du premier ordre définissant les estimations FIML peuvent donc s'écrire

$$\sum_{j=1}^g \hat{\sigma}^{ij} \hat{\mathbf{Z}}_i^\top(\mathbf{y}_j - \mathbf{Z}_j\hat{\boldsymbol{\delta}}_j) = \mathbf{0}, \quad \text{pour } i = 1, \dots, g. \quad (18.72)$$

Les conditions (18.72) apparaissent désormais sous une forme très comparable à celle des conditions (18.63) qui définissent l'estimateur 3SLS. En réalité, si nous notons  $\bar{\mathbf{Y}}_i$  la matrice de dimension  $n \times g_i$  des valeurs ajustées de la forme réduite *libre*, de sorte que  $\bar{\mathbf{Y}}_i = \mathbf{P}_X \mathbf{Y}_i$  for  $i = 1, \dots, g$ , alors

$$\mathbf{P}_X \mathbf{Z}_i = \mathbf{P}_X [\mathbf{X}_i \quad \mathbf{Y}_i] = [\mathbf{X}_i \quad \bar{\mathbf{Y}}_i] \equiv \bar{\mathbf{Z}}_i.$$

Ainsi la conditions (18.63) qui définit l'estimateur 3SLS peut s'écrire comme

$$\sum_{j=1}^g \tilde{\sigma}^{ij} \bar{\mathbf{Z}}_i^\top(\mathbf{y}_j - \mathbf{Z}_j\tilde{\boldsymbol{\delta}}_j) = \mathbf{0}. \quad (18.73)$$

Les différences existant entre les conditions qui définissent les estimations 3SLS et celles qui définissent les estimations FIML sont mises en évidence à partir de (18.73) et (18.72). Elles sont les suivantes:

- (i) l'estimation de la matrice de covariance provient des résidus 2SLS équation par équation en ce qui concerne les 3SLS, et des résidus FIML en ce qui concerne le FIML;

- (ii) Les valeurs ajustées de  $\mathbf{Y}$  employées en tant qu'instruments sont celles de la forme réduite non contrainte en ce qui concerne les 3SLS et celle du FIML en ce qui concerne le FIML.

Les deux différences reflètent le fait que, contrairement aux 3SLS, le FIML est une procédure d'estimation jointe: il faut résoudre simultanément les conditions (18.72) et les conditions (18.70) pour  $\Sigma$  si l'on veut obtenir une quelconque estimation ML.

Une autre façon d'établir la différence entre les deux procédures consiste à dire qu'elles emploient des estimations différentes des mêmes instruments optimaux. Ces instruments sont quelque peu délicats à écrire. Afin de le faire sans trop de difficulté, nous pouvons construire un vecteur de dimension  $ng$  constitué de toutes les contributions des moments empiriques. Sous forme partitionnée, ce vecteur peut s'écrire

$$[\mathbf{y}_1 - \mathbf{Z}_1\boldsymbol{\delta}_1 \vdots \cdots \vdots \mathbf{y}_g - \mathbf{Z}_g\boldsymbol{\delta}_g], \quad (18.74)$$

et un élément type est  $n$ -vector  $\mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\delta}_i$ . Au total, il faut identifier  $p \equiv \sum_{i=1}^g (g_i + k_i)$  paramètres, de sorte qu'il faut prémultiplier le vecteur (18.74) par exactement le nombre de vecteurs lignes, chacun étant de dimension  $ng$ , si l'on veut obtenir les équations définissantes pour ces estimations. On peut voir sans grande difficulté que la matrice de dimension  $p \times ng$  nécessaire à l'obtention de (18.72) ou de (18.73) est constituée de blocs de la forme  $\sigma^{ij}\mathbf{W}_i^\top$ , où  $\mathbf{W}_i$  indique une matrice de la forme  $[\mathbf{X}\boldsymbol{\Pi}_i \quad \mathbf{X}_i]$  pour un choix donné des matrices  $\boldsymbol{\Pi}_i$  de dimension  $n \times g_i$ . Ce bloc type est une matrice de dimension  $(g_i + k_i) \times n$ , ce qui est cohérent.

Les estimateurs 3SLS et FIML diffèrent selon la manière de choisir  $\Sigma$  et les matrices  $\boldsymbol{\Pi}_i$ . Les instruments optimaux réel, mais non observables, sont donnés en posant  $\Sigma$  égale à la véritable matrice de covariance des erreurs  $\Sigma_0$  et en posant  $\boldsymbol{\Pi}_i = \mathbf{B}_0\boldsymbol{\Gamma}_0^{-1}$ , à l'aide des véritables matrices de paramètres. A l'évidence, aussi bien  $\tilde{\Sigma}$  que  $\hat{\Sigma}$  convergent vers  $\Sigma_0$ . Identiquement, les matrices  $\tilde{\boldsymbol{\Pi}}$  telle que  $\tilde{\mathbf{Y}} = \mathbf{P}_X\mathbf{Y} = \mathbf{X}\tilde{\boldsymbol{\Pi}}$  obtenue de la forme réduite contrainte que la matrice  $\hat{\mathbf{B}}\hat{\boldsymbol{\Gamma}}^{-1}$  obtenue par l'estimation FIML convergent vers  $\mathbf{B}_0\boldsymbol{\Gamma}_0^{-1}$ . Les deux procédures emploient par conséquent des estimations convergentes des véritables instruments optimaux, de sorte que les deux sont asymptotiquement équivalentes et asymptotiquement efficaces. Remarquons que cette conclusion ne s'applique qu'à l'estimation de  $\boldsymbol{\Gamma}$  et  $\mathbf{B}$ : les procédures *ne sont pas* équivalentes en ce qui concerne l'estimation de la matrice de covariance  $\Sigma$ .

On peut obtenir l'équivalence numérique entre le FIML et les 3SLS en itérant ces derniers. A chaque itération, les résidus de la précédente étape sont utilisés pour générer les estimations actualisées de  $\Sigma$ , alors que les estimations paramétriques de la précédente étape sont utilisées pour générer les estimations actualisées de  $\boldsymbol{\Pi}$ . Une telle procédure itérative, dont l'intérêt reste surtout théorique, débute par les 3SLS et converge vers le FIML pour tous les paramètres, incluant ceux de  $\Sigma$ . Cette opération itérative, et de nombreuses autres, sont abordées par Hendry (1976), qui fournit également une

bibliographie exhaustive de la plupart des thèmes de la littérature consacrée aux équations simultanées existant à cette époque.

Comme nous l'avons suggéré lors de la Section 18.4, un moyen pratique de calculer une estimations de la matrice de covariance de l'estimateur FIML de  $\mathbf{F}$  et  $\mathbf{B}$  consiste à employer une expression comparable à (18.65). Si nous remplaçons l'estimation 3SLS  $\tilde{\Sigma}$  par l'estimation FIML  $\hat{\Sigma}$ , et les matrices  $\mathbf{P}_X \mathbf{Z}_i$  des 3SLS par les matrices  $\hat{\mathbf{Z}}_i$  du FIML, le résultat est

$$[\hat{\sigma}^{ij} \hat{\mathbf{Z}}_i^\top \hat{\mathbf{Z}}_j]^{-1}.$$

De même que le LIML appliqué à une équation est un cas dégénéré du FIML appliqué à ladite équation suridentifiée, les 2SLS sont un cas dégénéré des 3SLS appliqué à une équation suridentifiée unique d'un système global par ailleurs juste identifié. Ce résultat est d'une grande importance pratique, bien que la démonstration ne soit guère intéressante, et donc éludée. Le résultat implique que la raison invoquée dans la Section 18.5 qui nous conduit parfois à préférer le LIML au FIML, à savoir que cela évite d'imposer des contraintes de suridentification éventuellement inexactes, conduirait chaque expérimentateur dans un contexte de moindres carrés à ne jamais dépasser le stade des 2SLS. Compte tenu du fait que le surcroît de calcul pour obtenir les 3SLS par rapport aux 2SLS est considérable si l'on ne s'intéresse qu'à une seule équation, il est fondamental de réaliser que ce travail supplémentaire ne procure aucun avantage à moins que certaines équations du système ne soient suridentifiées.

Etant donné que les 3SLS sont un cas particulier de l'estimation par GMM, on peut les généraliser pour tenir compte d'une hétéroscédasticité de forme inconnue des aléas, chose impossible à réaliser avec le FIML. Si nous ne disposons d'aucune information quant à la forme de l'hétéroscédasticité, alors nous ne pouvons pas améliorer le choix (18.56) des conditions portant sur les moments empiriques employée pour l'identification des paramètres. Par contre nous pouvons remplacer l'estimation (18.58) de leur matrice de covariance basée sur l'hypothèse d'homoscédasticité par une estimation robuste à l'hétéroscédasticité. Avec des aléas corrélés en série, (18.57) reste une expression correcte pour la matrice de covariance des moments empiriques. Un bloc type de cette matrice est

$$\sum_{t=1}^n E(u_{ti} u_{tj} \mathbf{X}_t^\top \mathbf{X}_t).$$

Il est clair que, tout comme pour les autres HCCME, il est possible d'estimer de façon convergente  $1/n$  fois cette matrice par

$$\frac{1}{n} \sum_{t=1}^n E(\tilde{u}_{ti} \tilde{u}_{tj} \mathbf{X}_t^\top \mathbf{X}_t),$$

que l'on peut écrire plus simplement sous la forme

$$\frac{1}{n} \mathbf{X}^\top \tilde{\boldsymbol{\Omega}}_{ij} \mathbf{X} \quad (18.75)$$

si l'on pose la définition  $\tilde{\boldsymbol{\Omega}}_{ij} = \text{diag}(\tilde{u}_{ti}\tilde{u}_{tj})$ , pour  $i, j = 1, \dots, g$ . Si nous employons cette expression pour élaborer une fonction critère basée sur les conditions portant sur les moments empiriques (18.56), nous aboutissons à un nouvel estimateur, défini par les équations

$$\sum_{j=1}^g \mathbf{Z}_i^\top \mathbf{X} (\mathbf{X}^\top \tilde{\boldsymbol{\Omega}}_{ij} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y}_j - \mathbf{Z}_j \boldsymbol{\delta}_j) = \mathbf{0}.$$

La résolution de ces équations nous conduit à l'estimateur

$$\tilde{\boldsymbol{\delta}} = [\mathbf{Z}_i^\top \mathbf{X} (\mathbf{X}^\top \tilde{\boldsymbol{\Omega}}_{ij} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}_j]^{-1} \left[ \sum_{j=1}^g \mathbf{Z}_i^\top \mathbf{X} (\mathbf{X}^\top \tilde{\boldsymbol{\Omega}}_{ij} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}_j \right]. \quad (18.76)$$

Il n'est pas surprenant de retrouver en (18.76) une structure très comparable à celle de l'estimateur H2SLS (17.44), aussi l'appellerons-nous **estimateur H3SLS**. On peut estimer sa matrice de covariance asymptotique par l'inverse de la matrice avec le bloc type

$$\mathbf{Z}_i^\top \mathbf{X} (\mathbf{X}^\top \tilde{\boldsymbol{\Omega}}_{ij} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}_j.$$

En présence d'hétéroscédasticité de forme inconnue, l'estimateur H3SLS devrait être plus efficace, asymptotiquement que celui des 3SLS ou du FIML. Malgré tout, ses performances avec des échantillons finis sont pratiquement inconnus à ce jour.

Il est évident que nous pourrions généraliser l'estimateur H3SLS encore davantage à l'aide d'un estimateur HAC de la matrice de covariance à la place de la HCCME (18.75); consulter, par exemple, Gallant (1987, Chapitre 6). Cependant, c'est une stratégie adéquate tant que la présence de corrélation en série reste compatible avec le modèle correctement spécifié et que la taille d'échantillon est relativement importante. Pour la plupart des applications sur données chronologiques, le FIML ou les 3SLS restent les estimateurs systémiques préférés, du fait que l'hétéroscédasticité sera largement absente, alors que la corrélation en série largement répandue si le modèle est mal spécifié. Quoi qu'il en soit, lorsque la taille de l'échantillon est importante et que l'hétéroscédasticité se manifeste fortement, comme c'est le cas avec de nombreuses applications sur données en coupe transversale, il est fort probable que l'estimateur H3SLS soit l'estimateur systématique le plus approprié.

## 18.7 MODÈLES D'EQUATIONS SIMULTANÉES NON LINÉAIRES

A ce stade de l'exposé, nous avons très peu parlé des **modèles d'équations simultanées non linéaires**. Un modèle d'équations simultanées peut être non linéaire de trois manières possibles. Pour la première,  $\mathbf{Y}_t$  peut dépendre de fonctions non linéaires de quelques variables exogènes ou prédéterminées. Comme d'habitude, ce type de non linéarité n'engendre pas de problème et peut être géré de façon simple en redéfinissant  $\mathbf{X}_t$ . Pour la deuxième, certains paramètres peuvent agir de manière non linéaire dans le modèle structurel pour  $\mathbf{Y}_t$ , sans doute parce qu'ils sont soumis à des contraintes non linéaires. C'est le genre de non linéarité que nous avons traité fréquemment avec l'estimation de modèles de régression non linéaire, et elle ne cause pas de problème supplémentaire dans le contexte des modèles d'équations simultanées. Enfin, pour la troisième, il peut exister des non linéarités provoquées par les variables endogènes. Ce type de non linéarité ne pose pas non plus de problème sérieux supplémentaire.

Le problème avec les modèles qui sont non linéaires du fait des variables endogènes est que pour de tels modèles il n'existe aucun équivalent à la forme réduite non contrainte d'un modèle d'équations simultanées linéaire. Il est habituellement difficile voire impossible d'obtenir les variables endogènes en fonction de variables exogènes et des aléas. Même lorsque cela est possible,  $\mathbf{Y}_t$  dépendra presque toujours de façon non linéaire à la fois des exogènes et des aléas. Soit, par exemple, le modèle simple à deux équations

$$\begin{aligned} y_1 &= \alpha y_2 + \mathbf{X}_1 \boldsymbol{\beta}_1 + u_1 \\ y_2 &= \gamma_1 y_1 + \gamma_2 y_1^2 + \mathbf{X}_2 \boldsymbol{\beta}_2 + u_2, \end{aligned} \tag{18.77}$$

où la notation reste conventionnelle et où l'indice  $t$  a été supprimé pour ne pas surcharger les expressions. Si nous substituons le membre de droite de la première équation de (18.77) dans la seconde, nous obtenons

$$y_2 = \gamma_1 (\alpha y_2 + \mathbf{X}_1 \boldsymbol{\beta}_1 + u_1) + \gamma_2 (\alpha y_2 + \mathbf{X}_1 \boldsymbol{\beta}_1 + u_1)^2 + \mathbf{X}_2 \boldsymbol{\beta}_2 + u_2.$$

Puisque cette équation est une forme quadratique en  $y_2$ , elle possèdera habituellement deux solutions. Selon les valeurs paramétriques et les valeurs des  $\mathbf{X}_i$  et des  $u_i$ , les deux solutions peuvent être réelles ou pas. Même *s'il existe* une solution réelle, elle ne sera généralement pas linéaire en les variables exogènes. Par conséquent, le simple usage des composantes de  $\mathbf{X}_1$  et de  $\mathbf{X}_2$  en tant qu'instruments ne sera pas optimal.

Cet exemple illustre la nature des problèmes que l'on peut rencontrer avec tout modèle d'équations simultanées qui n'est pas linéaire en les variables endogènes. Nous sommes au moins confrontés à un problème de choix des instruments. Une approche, discutée dans la Section 7.6, consiste à employer des puissances et même des produits croisés des variables exogènes en



tant qu'instruments, en même temps que les variables exogènes elles-mêmes. Si la taille de l'échantillon est suffisamment importante, cette approche est judicieuse, mais dans de nombreux cas il sera difficile de déterminer le nombre d'instruments à employer, et même de savoir lesquels employer. L'ajout d'instruments améliorera généralement l'efficacité asymptotique mais tendra également à accroître le biais avec des échantillons finis. Plus sérieusement, il est fort possible d'estimer un modèle qui ne peut pas être résolu pour des valeurs tout à fait raisonnables des variables exogènes et des aléas. Ainsi il faudrait probablement éviter d'employer des modèles qui sont non linéaires en les variables endogènes, si cela est possible.

Il semble que le LIML ne soit pas une procédure viable pour l'estimation de modèles d'équations simultanées non linéaires. La procédure LIML classique discutée dans la Section 18.5 est conçue exclusivement pour les modèles linéaires. On peut imaginer obtenir des estimations LIML d'une équation structurelle non linéaire en employant un programme pour le FIML non linéaire appliqué à un système constitué d'une seule équation structurelle et de  $g - 1$  équations linéaires sous forme réduite. Cela ne serait cohérent que si les équations sous forme réduite étaient en fait linéaires, ce qui ne sera presque jamais le cas. Ainsi, pour l'estimation d'équations isolées, les seules procédures adéquates sont celles basées sur les variables instrumentales.

Nous avons discuté de l'estimation de modèles non linéaires constitués d'une seule équation par les méthodes IV dans la Section 7.6, et il reste seulement quelques compléments à livrer sur ce sujet. Supposons que l'équation structurelle qui nous intéresse puisse s'écrire

$$\mathbf{y} = \mathbf{x}(\boldsymbol{\delta}) + \mathbf{u},$$

où  $\boldsymbol{\delta}$  est un vecteur composé de  $l$  paramètres, et le vecteur de fonctions non linéaires  $\mathbf{x}(\boldsymbol{\delta})$  dépend implicitement d'au moins une variable endogène et d'un certain nombre de variables exogènes et prédéterminées. Alors si  $\mathbf{W}$  désigne une matrice d'instruments de dimension  $n \times m$ , nous avons vu que les estimations IV peuvent être calculées en minimisant la fonction critère

$$(\mathbf{y} - \mathbf{x}(\boldsymbol{\delta}))^\top \mathbf{P}_W (\mathbf{y} - \mathbf{x}(\boldsymbol{\delta})). \quad (18.78)$$

Les estimations qui en résultent sont souvent nommées **moindres carrés non linéaires en deux étapes** ou estimations **NL2SLS**, si l'on se réfère à la terminologie d'Amemiya (1974), bien que ces estimations ne soient pas obtenues en deux étapes. Nous avons vu ce détail dans la Section 7.6.

La fonction critère (18.78) peut se dériver comme une procédure GMM en débutant par les conditions portant sur les moments

$$E(\mathbf{W}^\top (\mathbf{y} - \mathbf{x}(\boldsymbol{\delta}))) = \mathbf{0}$$

et en supposant que  $E(\mathbf{u}\mathbf{u}^\top) = \sigma^2 \mathbf{I}$ . Cette hypothèse peut se révéler parfois trop contraignante. Si elle était correcte, la minimisation de (18.78) produirait

des estimations non efficaces et une estimation non convergentes de la matrice de covariance des paramètres estimés. Une hypothèse plus souple est que  $E(\mathbf{u}\mathbf{u}^\top) = \mathbf{\Delta}$ , où  $\mathbf{\Delta}$  est une matrice diagonale dont les éléments diagonaux sont inconnus (mais finis). Nous pouvons obtenir des estimations analogues aux estimations H2SLS de la Section 17.3 à l'aide d'une procédure en deux étapes. Dans la première étape, nous minimisons (18.78), de manière à obtenir des estimations paramétriques convergentes mais non efficaces et des résidus  $\tilde{u}_t$ , et nous utilisons ces derniers pour construire la matrice  $\mathbf{W}^\top \tilde{\mathbf{\Delta}} \mathbf{W}$ , où  $\tilde{\mathbf{\Delta}}$  a comme élément type  $\tilde{u}_t^2$ . Dans la seconde étape, nous minimisons la fonction critère

$$(\mathbf{y} - \mathbf{x}(\boldsymbol{\delta}))^\top \mathbf{W} (\mathbf{W}^\top \tilde{\mathbf{\Delta}} \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{x}(\boldsymbol{\delta})).$$

Comme d'habitude, nous pourrions abandonner l'hypothèse de diagonalité de  $\mathbf{\Delta}$  et employer un estimateur HAC, si cela s'avérait utile (voir les remarques à la fin de la section précédente).

L'estimation systématique des modèles d'équations simultanées non linéaires relève typiquement d'une sorte de procédure IV (ou GMM) ou FIML. Nous discuterons brièvement de ces deux approches à tour de rôle. Supposons que la  $i^{\text{ième}}$  équation du système puisse s'écrire pour toutes les observations sous la forme

$$\mathbf{f}_i(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) = \mathbf{u}_i, \quad (18.79)$$

où  $\mathbf{f}_i(\cdot)$  est un vecteur de dimension  $n$  de fonctions non linéaires,  $\mathbf{u}_i$  est un vecteur de dimension  $n$  d'aléas, et où  $\boldsymbol{\theta}$  est un vecteur de dimension  $p$  de paramètres qu'il s'agit d'estimer. En général, toutes les variables endogènes et exogènes et tous les paramètres peuvent apparaître dans n'importe quelle équation, compte tenu des contraintes quelconques que l'on peut vouloir leur imposer pour identifier le système.

La première étape dans toute procédure IV consiste à choisir les instruments que l'on va utiliser. Si le modèle est non linéaire seulement en les paramètres, la matrice des instruments optimaux est  $\mathbf{X}$ . Cependant, comme nous l'avons vu, il n'existe pas de moyen simple de choisir les instruments pour les modèles qui sont non linéaires en une ou plusieurs variables endogènes. La théorie de la Section 17.4 peut s'appliquer, bien entendu, mais le résultat qu'elle entraîne n'est pas d'un grand intérêt pratique. Il apparaît que sous les hypothèses habituelles sur les termes d'erreur, à savoir leur homoscedasticité et leur indépendance en série mais pas entre les équations, la matrice des instruments  $\mathbf{W}$  sera optimale si  $\mathcal{S}(\mathbf{W})$  correspond à l'union des sous-espaces engendrés par les colonnes de  $E(\partial \mathbf{f}_i / \partial \boldsymbol{\theta})$ . Ce résultat est dû à Amemiya (1977). Il reste pertinent mais généralement, il n'est pas utile dans la pratique. Pour l'instant, nous supposons simplement qu'une *certaine* matrice d'instruments  $\mathbf{W}$  de dimension  $n \times m$  est disponible, avec  $m \geq p$ .

Une procédure IV non linéaire pour l'estimation systématique, comparable dans l'esprit à la procédure équation par équation des NL2SLS basée sur la minimisation de (18.78), fut proposée à l'origine par Jorgenson et Laffont (1974) et fut nommée **moindres carrés en trois étapes**, ou **NL3SLS**.

L'appellation est quelque peu trompeuse, pour une raison identique à celle qui fait que le nom "NL2SLS" est également trompeuse. Par analogie avec (18.60), la fonction critère que nous voudrions réellement minimiser est

$$\sum_{i=1}^g \sum_{j=1}^g \sigma^{ij} \mathbf{f}_i^\top(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \mathbf{P}_W \mathbf{f}_j(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}). \quad (18.80)$$

Quoi qu'il en soit, dans la pratique, les éléments  $\sigma^{ij}$  de l'inverse de la matrice de covariance contemporaine  $\boldsymbol{\Sigma}$  ne seront pas connus et il nous faudra les estimer. Plusieurs possibilités s'offrent à nous. On peut tout d'abord employer les NL2SLS pour chaque équation séparément. Cela sera traditionnellement plus aisé, mais pas toujours possible si certains paramètres ne sont identifiés que grâce à des contraintes d'équations croisées. Une autre approche qui fonctionnera dans ce cas consiste à minimiser la fonction critère

$$\sum_{i=1}^g \sum_{j=1}^g \mathbf{f}_i^\top(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \mathbf{P}_W \mathbf{f}_j(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}), \quad (18.81)$$

pour laquelle la matrice de covariance  $\boldsymbol{\Sigma}$  est remplacée par la matrice identité. La minimisation de (18.81) conduira à un estimateur qui sera à l'évidence un estimateur GMM valable, et par conséquent convergent même s'il n'est pas efficace. Quel que soit l'estimateur non efficace utilisé à l'étape initiale, il produira  $g$  vecteur de résidus  $\hat{\mathbf{u}}_i$  à partir desquels on peut estimer de façon convergente la matrice  $\boldsymbol{\Sigma}$ , exactement de la même manière que pour les modèles linéaires; voir (18.62). On obtient alors la fonction critère

$$\sum_{i=1}^g \sum_{j=1}^g \hat{\sigma}^{ij} \mathbf{f}_i^\top(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \mathbf{P}_W \mathbf{f}_j(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}), \quad (18.82)$$

en remplaçant les  $\sigma^{ij}$  inconnus dans (18.80) par les éléments  $\hat{\sigma}^{ij}$  de l'inverse de l'estimation de  $\boldsymbol{\Sigma}$ . Cette fonction critère peut véritablement être minimisée dans la pratique.

Comme d'habitude, la valeur minimisée de la fonction critère (18.82) fournit une statistique de test pour les contraintes de suridentification; voir les Sections 7.8 et 17.6. Si le modèle et les instruments sont correctement spécifiés, cette statistique de test sera asymptotiquement distribuée suivant une  $\chi^2(m-p)$ ; souvenons-nous que les instruments sont au nombre de  $m$  et que les paramètres libres sont au nombre de  $p$ . De plus, si le modèle est estimé sans contrainte puis sous  $r$  contraintes distinctes, la différence entre les deux valeurs des fonctions critères aura une distributions asymptotique du  $\chi^2(r)$ . Si cette dernière statistique de test doit être utilisée, il est fondamental que la même estimation de  $\boldsymbol{\Sigma}$  soit employée dans les deux estimations, car autrement la statistique de test peut même ne pas être positive avec des échantillons finis.

Lorsque la taille de l'échantillon est importante, il est peut être plus facile d'obtenir des estimations efficaces en une étape plutôt que de minimiser (18.82). Supposons que l'on note  $\hat{\theta}$  les estimations efficaces initiales, qui peuvent être soit des estimations NL2SLS soit des estimations systématiques basées sur (18.81). Un développement en série de Taylor de  $f_i(\theta) \equiv f_i(Y, X, \theta)$  autour de  $\hat{\theta}$  est

$$f_i(\hat{\theta}) + F_i(\hat{\theta})(\theta - \hat{\theta}),$$

où  $F_i$  est une matrice de dimension  $n \times p$  des dérivées de  $f_i(\theta)$  par rapport aux  $p$  éléments de  $\theta$ . Si quelques paramètres n'apparaissent pas dans l'équation  $i$ , les colonnes correspondantes de  $F_i$  seront identiquement nulles. Les estimations en une étape, qui seront asymptotiquement équivalentes aux estimations NL3SLS, sont simplement  $\hat{\theta} = \hat{\theta} - \hat{t}$ , où  $\hat{t}$  désigne le vecteur des estimations 3SLS *linéaires*

$$\hat{t} = [\hat{\sigma}^{ij} \hat{F}_i^\top P_W \hat{F}_j]^{-1} \left[ \sum_{j=1}^g \hat{\sigma}^{ij} \hat{F}_i^\top P_W \hat{f}_j \right]. \quad (18.83)$$

Cette expression doit être comparée à (18.64).

Il est clair que l'on peut généraliser les NL3SLS pour gérer une hétéroscédasticité de forme inconnue, une corrélation sérielle de forme inconnue, ou les deux simultanément. Par exemple, afin de tenir compte d'une hétéroscédasticité, nous remplacerions simplement la matrice  $P_W$  dans (18.82) et (18.83) par la matrice

$$W(W^\top \hat{\Omega}_{ij} W)^{-1} W^\top,$$

où, par analogie avec (18.76),  $\hat{\Omega}_{ij} = \text{diag}(\hat{u}_{ti} \hat{u}_{tj})$  pour  $i, j = 1, \dots, g$ . Les estimations initiales  $\hat{\theta}$  peuvent ne pas tenir compte de l'hétéroscédasticité. Pour une discussion plus détaillée sur cette sorte de procédure, et de NL3SLS en général, consulter Gallant (1987, Chapitre 6).

L'autre méthode d'estimation systématique qui est largement employée est celle du **FIML non linéaire**. Pour l'examiner, il est judicieux d'écrire le système d'équations à estimer non pas sous la forme (18.79) mais plutôt sous la forme

$$h_t(Y_t, X_t, \theta) = U_t, \quad U_t \sim \text{NID}(0, \Sigma), \quad (18.84)$$

où  $\theta$  demeure un vecteur de  $p$  paramètres,  $h_t$  un vecteur de dimension  $1 \times g$  de fonctions non linéaires, et  $U_t$  un vecteur de dimension  $1 \times g$  de termes d'erreur. Pour admettre que (18.79) et (18.84) sont de formes comparables il suffit d'imaginer que le  $i^{\text{ième}}$  élément de  $h_t(\cdot)$  est identique au  $t^{\text{ième}}$  élément de  $f_i(\cdot)$ .

La densité du vecteur  $U_t$  est

$$(2\pi)^{-g/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} U_t \Sigma^{-1} U_t^\top\right).$$

Pour se ramener à la densité de  $\mathbf{Y}_t$ , nous devons remplacer  $\mathbf{U}_t$  par  $\mathbf{h}_t(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta})$  et multiplier par le terme jacobien  $|\det \mathbf{J}_t|$ , où  $\mathbf{J}_t \equiv \partial \mathbf{h}_t(\boldsymbol{\theta}) / \partial \mathbf{Y}_t$ , c'est-à-dire la matrice de dimension  $g \times g$  des dérivées de  $\mathbf{h}_t$  par rapport aux éléments de  $\mathbf{Y}_t$ . La résultat est

$$(2\pi)^{-g/2} |\det \mathbf{J}_t| |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{h}_t(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1} \mathbf{h}_t^\top(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta})\right).$$

Il s'ensuit immédiatement que la fonction de logvraisemblance est

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = & -\frac{ng}{2} \log(2\pi) + \sum_{t=1}^n \log |\det \mathbf{J}_t| - \frac{n}{2} \log |\boldsymbol{\Sigma}| \\ & - \frac{1}{2} \sum_{t=1}^n \mathbf{h}_t(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta}) \boldsymbol{\Sigma}^{-1} \mathbf{h}_t^\top(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta}). \end{aligned} \quad (18.85)$$

Cette expression peut être maximisée par rapport à  $\boldsymbol{\Sigma}$  et le résultat injecté pour mener à la fonction de logvraisemblance concentrée

$$\begin{aligned} \ell^c(\boldsymbol{\theta}) = & -\frac{ng}{2} (\log(2\pi) + 1) + \sum_{t=1}^n \log |\det \mathbf{J}_t| \\ & - \frac{n}{2} \log \left| \frac{1}{n} \sum_{t=1}^n \mathbf{h}_t^\top(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta}) \mathbf{h}_t(\mathbf{Y}_t, \mathbf{X}_t, \boldsymbol{\theta}) \right|. \end{aligned} \quad (18.86)$$

De toute évidence, il existe une forte ressemblance entre (18.85) et (18.86) et leurs contreparties (18.28) et (18.30) pour le cas linéaire. La différence majeure est que le terme jacobien dans (18.85) et (18.86) correspond à la somme des logarithmes de  $n$  déterminants différents. Ainsi à chaque évaluation de ces fonctions de logvraisemblance, il faut calculer  $n$  déterminants différents. Cela peut s'avérer coûteux lorsque  $g$  ou  $n$  est important. Bien sûr, le problème disparaît si le modèle est linéaire n les variables endogènes, puisqu'alors  $\mathbf{J}_t$  sera constant.

Une difficulté avec le FIML non linéaire est que l'on ne sait pas trop bien comment tester les contraintes de suridentification, ni même à quoi elles peuvent ressembler dans de nombreux cas. Dans le contexte d'un modèle d'équations simultanées linéaire, toute forme structurelle impose des contraintes non linéaires à la forme réduite non contrainte, et un test LR permet de tester simplement ces contraintes. Cependant, dans le cas d'un modèle d'équations simultanées non linéaire en les variables endogènes, nous ne pouvons en général pas même écrire la FRL, let alone estimate it. On peut toujours tester n'importe quelle contrainte à l'aide des tests classiques, qu'il s'agisse de contraintes d'équations croisées ou de contraintes portant sur une équation isolée. Mais il sera en général impossible de tester toutes les contraintes de suridentification en même temps. Il existe un problème connexe

avec l'estimation NL3SLS, bien sûr. Bien que la valeur minimisée de la fonction critère (18.82) fournisse une statistique de test, elle ne sera valable que pour les contraintes de suridentification associées à une matrice d'instruments particulière  $\mathbf{W}$ , qui peut parfaitement ne pas procurer une approximation satisfaisante à la véritable forme réduite non contrainte, qui est inconnue.

La relation entre le FIML non linéaire et les NL3SLS n'est pas de nature comparable à celle qui existe entre le FIML linéaire et les 3SLS. Les deux méthodes non linéaires seront asymptotiquement équivalentes lorsque le modèle est linéaire en les variables endogènes. Toutefois, dans la majorité des situations, elles ne le seront pas. Dans l'éventualité d'une non équivalence, le FIML non linéaire sera plus efficace, asymptotiquement, que les NL3SLS. Mais cette plus grande efficacité se paye. Lorsque le FIML non linéaire et les NL3SLS ne sont pas équivalents, le premier peut être non convergent si les aléas sont en réalité distribués autrement que suivant la loi normale multivariée. Au contraire, comme nous l'avons vu, l'hypothèse de normalité n'est pas nécessaire pour assurer la convergence du FIML linéaire. Pour plus de détails sur ces points, consulter Amemiya (1977) et Phillips (1982). Amemiya (1985, Chapitre 8) et Gallant (1987, Chapitre 6) donnent des traitements plus explicites du FIML non linéaire que le notre.

Il existe une littérature véritablement vaste sur le calcul des estimations par le FIML non linéaire. Comme d'habitude, on peut employer de nombreux algorithmes différents pour maximiser la fonction de logvraisemblance et la fonction de logvraisemblance concentrée, dont certains exploitent des caractéristiques spéciales des classes particulières de modèles. Les références classiques sont Eisenpress et Greenstadt (1966), Chow (1973), Dagenais (1978), Belsley (1979, 1980), Fair and Parke (1980), Parke (1982), et Quandt (1983).

## 18.8 CONCLUSION

Le fait que nous traitions un thème aussi important que les modèles d'équations simultanées aussi tard peut heurter certains lecteurs. Nous avons bien évidemment abordé certains aspects du problème dans le Chapitre 7, en tant que contribution à notre traitement des variables instrumentales. La raison de ce retard volontaire est que nous voulions que le lecteur ait acquis une compréhension claire de l'estimation et des tests de spécification par maximum de vraisemblance et de la méthode des moments généralisée. Cela nous a alors permis de développer toutes les méthodes d'estimation et de test discutées dans ce chapitre en tant qu'applications immédiates du ML et de la GMM. Si l'on admet cela, il est beaucoup plus facile de comprendre les modèles d'équations simultanées et les techniques statistiques qui leur sont associées.

## TERMES ET CONCEPTS

causalité au sens de Granger	maximum de vraisemblance en
condition d'ordre pour l'identification	information limitée (LIML)
condition de rang pour l'identification	modèles d'équations simultanées
contraintes d'équation croisées	modèles d'équations simultanées
contraintes de suridentification	linéaire
doubles moindres carrés non linéaires	modèles d'équations simultanées non
(NL2SLS)	linéaire
estimateur de classe $K$	non causalité au sens de Granger
estimateur du ratio de moindre	paramètres de nuisance
variance	paramètre d'intérêt
estimateur H3SLS	super exogénéité
exogénéité	système récursif
exogénéité faible	triples moindres carrés (3SLS)
exogénéité stricte	triples moindres carrés non linéaires
FIML non linéaire	(NL3SLS)
fonction de logvraisemblance partielle	variable endogène
forme réduite contrainte (FRC)	variable exclue
forme réduite libre (FRL)	variable exogène
maximum de vraisemblance en	variable incluse
information complète (FIML)	variable prédéterminée

# Chapitre 19

## Modèles de Régressions pour Données Chronologiques

### 19.1 INTRODUCTION

Un nombre conséquent d'études économétriques appliquées utilisent des données chronologiques, et nombreux sont les problèmes économétriques qui sont liés au seul usage de ce genre de données. L'un d'entre eux est la corrélation en série, dont nous avons largement parlé au cours du Chapitre 10. Dans ce chapitre et celui qui suit, nous discuterons d'autres problèmes que l'on rencontre fréquemment lorsque l'on utilise les données chronologiques ou des méthodes susceptibles de les traiter. Dans la Section 19.2, nous aborderons le problème des régressions "erronées" entre des séries économiques temporelles. Cette section introduit quelques concepts importants qui feront l'objet du Chapitre 20, lorsque nous parlerons des racines unitaires et de la cointégration. La Section 19.3 traite l'estimation des retards échelonnés. La Section 19.4 concerne les modèles de régression dynamique, dans lesquels un ou plusieurs retards de la variable dépendante apparaissent dans les régresseurs. Nous discuterons de l'estimation des modèles à vecteur autorégressif pour des séries chronologiques multivariées dans la Section 19.5. Les deux sections finales traitent de la saisonnalité. La Section 19.6 fournit une introduction aux procédures d'ajustement saisonnier, et la Section 19.7 discute des moyens variés de modéliser les variations saisonnières dans les modèles de régression.

### 19.2 RÉGRESSIONS ERRONÉES

De nombreuses séries temporelles économiques ont une tendance croissante dans le temps. Cette observation est sans doute vraie pour la plupart des séries qui mesurent, ou qui sont mesurées avec les prix nominaux, du moins pour notre siècle. Elle est également vraie pour des données chronologiques qui mesurent les niveaux des variables économiques réelles, telles que la consommation, la production, l'investissement, les importations et les exportations. De nombreuses séries tendanciellées peuvent être généralement caractérisées



par l'un des deux modèles suivants:

$$y_t = \gamma_1 + \gamma_2 t + u_t \quad \text{et} \quad (19.01)$$

$$y_t = \delta_1 + y_{t-1} + u_t, \quad (19.02)$$

où les aléas  $u_t$  ne seront, en général, ni indépendants ni identiquement distribués. Ils seront cependant stationnaires si le modèle est bien adapté à la série temporelle concernée. Le premier modèle, (19.01), indique que  $y_t$  est **stationnaire en tendance**, c'est-à-dire qu'il est stationnaire autour d'une tendance. Par contraste, le second modèle, (19.02), indique paramètre de dérive  $\delta_1$  dans (19.02) joue un rôle comparable au paramètre de tendance  $\gamma_2$  dans (19.01), puisque les deux donnent une orientation croissante à  $y_t$  à travers le temps. Mais le comportement de  $y_t$  est très différent dans les deux cas, parce qu'enlever la tendance de  $y_t$  dans le premier cas en fait une variable stationnaire, alors que dans le second cas, ce n'est pas exact.

Il existe une littérature importante consacrée à la détermination du modèle qui caractérise le mieux la plupart des séries temporelles, détermination qui arbitre entre le modèle stationnaire en tendance (19.01) et le modèle à marche aléatoire avec dérive (19.02). L'article de Nelson et Plosser (1982) est une référence classique, celui de Campbell et Mankiw (1987) est plus récent, et celui de Stock et Watson (1988a) offre une discussion excellente de nombreux résultats de nombreux résultats. Dans le prochain chapitre nous discuterons des méthodes que l'on peut employer pour savoir par lequel de ces modèles une série temporelle donnée est le mieux caractérisée. Pour l'instant, ce qui nous préoccupe est ce qui survient si l'on utilise des séries chronologiques, qui sont décrites par l'un ou l'autre de ces modèles, en tant que variables dépendantes ou indépendantes dans un modèle de régression.

Si une série chronologique dont l'élément type est  $x_t$  est toujours croissante, alors  $n^{-1} \sum_{t=1}^n x_t^2$  divergera vers  $+\infty$ . Ainsi, si l'on utilise une telle série en tant que régresseur dans un modèle de régression linéaire, la matrice  $n^{-1} \mathbf{X}^\top \mathbf{X}$  ne peut pas tendre vers une matrice finie, définie positive. Toute la théorie asymptotique que nous avons utilisée dans cet ouvrage est donc inadaptée aux modèles pour lesquels n'importe quel régresseur est caractérisé par (19.01) ou par (19.02).<sup>1</sup> Cela ne signifie pas qu'il ne faut *jamaïs* poser une variables

<sup>1</sup> Le fait que la théorie asymptotique *standard* soit inadaptée à de tels modèles ne signifie pas qu'aucune théorie ne leur soit pas applicable. Par exemple, nous avons étudié un modèle simple de régression sur une tendance linéaire dans la Section 4.4 et nous avons conclu que l'estimateur des moindres carrés du coefficient du terme de tendance était convergent, mais avec une variance  $O(n^{-3})$  au lieu d'être  $O(n^{-1})$ . De plus, puisqu'il existe des TLC qui s'appliquent à de tels modèles, les procédures habituelles pour l'inférence sont asymptotiquement valables. Par exemple, si  $u_t \sim \text{IID}(0, \sigma^2)$  et  $S_n \equiv n^{-3/2} \sum_{t=1}^n t u_t$ , alors  $S_n$  a une distribution qui tend vers  $N(0, \sigma^2/3)$ . Remarquons que le facteur de normalisation ici est  $n^{-3/2}$  plutôt que  $n^{-1/2}$ .

de tendance dans le membre de droite d'une régression linéaire ou non linéaire. Puisque les échantillons observés sont finis, et parfois assez restreints, nous ne pouvons jamais assurer que la tendance est toujours croissante. De plus, les propriétés agréables, avec des échantillons finis, de la régression par moindres carrés sont maintenues que des régresserurs aient une tendance croissante ou pas. Mais si l'on veut s'appuyer sur la théorie asymptotique conventionnelle, il semblerait que la spécification de nos modèles sans variable à tendance affirmé dans le membre de droite soit une attitude prudente. Cela implique en retour que la variable dépendante ne peut pas avoir de tendance affirmée. L'approche la plus commune consiste à prendre les différences de toutes les variables avant de spécifier le modèle.

Une raison irrésistible qui motive la considération des différences premières est le phénomène de **régression erronée**. Il devrait être clair que si deux variables, disons  $y_t$  et  $x_t$ , toutes deux à tendance croissante, une régression de  $y_t$  sur  $x_t$  a de fortes chances de trouver une relation "significative" entre elles, même si la seule chose qu'elles ont en commun est cette tendance croissante. En réalité, le  $R^2$  pour une régression de  $y_t$  sur  $x_t$  et une constante tendra vers 1 alors que  $n \rightarrow \infty$  lorsque les deux séries peuvent être caractérisées par (19.01), même s'il n'y a pas de corrélation en série entre les deux parties aléatoires de  $y_t$  et de  $x_t$ . Les lecteurs trouveraient sans doute révélatrice la démonstration de ce résultat, et nous leur conseillons de consulter la Section 4.4 pour quelques résultats utiles.

Il est intuitivement très plausible que nous devrions observer des relations en apparence significatives, mais en réalité fausses, entre des variables sans lien mais à tendance croissante dans le temps. Granger et Newbold (1974) ont découvert ce qui semble être au premier abord une forme encore plus surprenante de régression erronée. Ils considèrent des séries temporelles générées par une **marche aléatoire sans dérive**, c'est-à-dire des séries générées par un processus comme  $y_t = y_{t-1} + u_t$ . Leur résultat, obtenu par des expériences Monte Carlo, est que si  $x_t$  and  $y_t$  sont des variables aléatoires indépendantes, le  $t$  de Student de  $\beta = 0$  dans la régression

$$y_t = \alpha + \beta x_t + u_t \quad (19.03)$$

rejette l'hypothèse nulle beaucoup plus souvent qu'il ne le devrait et tend à la rejeter d'autant plus souvent que la taille de l'échantillon,  $n$ , augmente. Ultérieurement, Phillips (1986) démontrera que ce  $t$  de Student rejettera constamment l'hypothèse nulle, asymptotiquement.

Quelques résultats Monte Carlo sur les régressions erronées figurent dans le Tableau 19.1. Chaque colonne décrit la proportion des fois, dans plus de 10,000 exécutions, où le  $t$  de Student de  $\beta = 0$  rejettera l'hypothèse nulle au niveau 5% dans une régression quelconque. Pour la colonne 1, la régression est (19.03) et à la fois  $x_t$  et  $y_t$  sont générées par des marches aléatoires indépendantes à aléas n.i.d. Pour la colonne 2,  $x_t$  et  $y_t$  sont identiques à celles de la première colonne, mais une variable dépendante retardée a été

**Tableau 19.1** Rejets Erronés et Taille d'Echantillon

$n$	Marche Aléatoire	Retard Ajouté	Dérive	Tendance
25	0.530	0.146	0.645	0.066
50	0.662	0.154	0.825	0.431
75	0.723	0.162	0.905	0.987
100	0.760	0.162	0.945	1.000
250	0.847	0.169	0.997	1.000
500	0.890	0.167	1.000	1.000
750	0.916	0.170	1.000	1.000
1000	0.928	0.169	1.000	1.000
2000	0.947	0.168	1.000	1.000

ajoutée à la régression. Pour les colonnes 3 et 4, la régression est simplement (19.03) à nouveau. Pour la troisième colonne,  $x_t$  et  $y_t$  sont toutes deux générées par des marches aléatoires avec dérive, le paramètre de dérive  $\delta_1$  étant égal à un cinquième de la valeur de l'écart type  $\sigma$  (ce rapport est le seul paramètre qui affecte la distribution du  $t$  de Student). Pour la colonne 4,  $x_t$  et  $y_t$  sont stationnaires en tendance, avec un coefficient de tendance  $\gamma_2$  égal à  $1/25$  de la taille de  $\sigma$ .

Les résultats dans les colonnes 3 et 4 de tableau ne sont guère surprenants, puisque  $x_t$  et  $y_t$  sont croissants. Le seul élément intéressant concernant ces résultats est la rapidité d'accroissement du nombre de rejets en fonction de la taille de l'échantillon. C'est une conséquence du fait que, dans ces deux cas, la masse d'information contenue dans l'échantillon augmente à un taux plus fort que  $n$ . Elle augmente bien sûr encore plus vite dans le cas d'une tendance que dans le cas d'un e marche aléatoire avec dérive.

Par contre, les résultats des colonnes 1 et 2 du tableau peuvent surprendre. Après tout,  $x_t$  et  $y_t$  sont des éries totalement indépendantes, et aucune ne contient de tendance. Alors pour quelle raison découvrons-nous souvent — très souvent en fait pour des tailles d'échantillon importantes — l'évidence d'une relation lorsque nous régressons  $y_t$  sur  $x_t$ ? Une réponse devrait être évidente après la lecture du Chapitre 12. Les  $t$  de Student significatifs ne nous indiquent pas que  $\beta \neq 0$  dans (19.03), puisque c'est en réalité un modèle incorrect. Ils nous indiquent simplement que l'hypothèse nulle, qui est (19.03) avec  $\beta = 0$ , est fausse. Elle est fausse parce que si  $y_t$  est générée par une marche aléatoire, alors  $y_t$  n'est pas égal à une constante plus un terme aléatoire stationnaire. Ainsi, lorsque nous testons l'hypothèse nulle, même contre une hypothèse alternative qui est également fausse, nous la rejetons souvent.

Cette justification intuitive n'est pas entièrement satisfaisante, quoi qu'il en soit. L'analyse asymptotique standard ne s'applique pas ici, car si  $y_t$  est générée par une marche aléatoire  $n^{-1} \sum_{t=1}^n y_t^2$  diverge. Par conséquent, l'analyse du Chapitre 12 n'est pas appropriée. De plus, l'explication intuitive

n'indique pas pourquoi, pour des tailles d'échantillon suffisamment importantes, une relation entre  $y_t$  et  $x_t$  apparaît toujours. On peut imaginer que puisque les processus qui génèrent  $x_t$  et  $y_t$  sont indépendants, toute corrélation entre les deux doit disparaître asymptotiquement, mais ce n'est pas le cas ici. L'explication de ces résultats nécessite une analyse asymptotique non standard d'un genre que nous verrons dans le prochain chapitre. Une référence classique est Phillips (1986) et l'article de Durlauf et Phillips (1988) offre des résultats plus approfondis.

Le fait que (19.03) soit un modèle mal spécifié n'est pas la seule clé du problème, ainsi que le montre la colonne 2. Ces résultats sont relatifs au modèle

$$y_t = \delta_1 + \beta x_t + \delta_2 y_{t-1} + u_t,$$

qui comprend le DGP en tant que cas particulier lorsque  $\delta_2 = 1$  et que les deux autres paramètres sont nuls. Malgré tout, l'hypothèse nulle  $\beta = 0$  est rejetée environ trois fois plus souvent qu'elle ne devrait l'être, et il n'y a rien que montre que cette tendance au rejet quasi systématique décline lorsque la taille d'échantillon  $n$  s'accroît. Le  $t$  de Student provoque un rejet excessif dans ce cas parce qu'il *n'est pas* asymptotiquement comme une  $N(0, 1)$ . Puisque les deux régresseurs sont ici générés par des marches aléatoires, la matrice  $n^{-1}\mathbf{X}^\top\mathbf{X}$  n'est pas finie définie positive, et la théorie asymptotiques standard ne s'applique plus. Comme nous allons le voir dans le prochain chapitre, il existe de nombreux cas comparables, pour lesquels les  $t$  de Student suivent des distributions non standard asymptotiquement. Ces distributions sont pour l'instant calculées généralement au moyen d'expériences Monte Carlo.

Une série qui suit une marche aléatoire, avec ou sans dérive, est souvent qualifiée d'**intégrée à l'ordre un**, ou  **$I(1)$**  pour aller vite. L'idée sur laquelle repose cette terminologie est qu'une série doit être différenciée une fois pour être stationnaire. Ainsi une série stationnaire est dite  **$I(0)$** . En principe, une série pourrait être intégrée à d'autres ordres. Il est possible de rencontrer occasionnellement une série  $I(2)$ , et si l'on différencie malencontreusement une série  $I(0)$ , le résultat est une série  $I(-1)$ . Néanmoins, la grande majorité des travaux économétriques appliqués traite des séries temporelles qui sont soit  $I(0)$  ou  $I(1)$ . Si une série est à l'origine  $I(1)$ , il est possible de la différencier une fois pour la rendre  $I(0)$ . Savoir quand il est nécessaire de différencier une série sera l'objet du prochain chapitre.

Dans le reste de ce chapitre, nous ferons l'hypothèse que toutes les séries sont  $I(0)$  et ne contiennent aucune tendance non stochastique. Ces h2 garantissent que ni une régression erronée ni des résultats asymptotiques non standards ne poseront problème. Ces h2 peuvent paraître malgré tout un vœu pieux. Par chance, les techniques dont nous discuterons dans le prochain chapitre rendent possible la garantie que ces h2 ne sont pas trop remises en cause dans la pratique.

### 19.3 RETARDS ÉCHELONNÉS

On pense souvent qu'une variable dépendante  $y_t$  doit dépendre de nombreuses valeurs actuelle et retardées d'une variable indépendantes  $x_t$ . Une modélisation de ce genre consiste à utiliser un **modèle à retards échelonnés** tel que

$$y_t = \alpha + \sum_{j=0}^q \beta_j x_{t-j} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \quad (19.04)$$

où il s'agit d'estimer la constante  $\alpha$  et les coefficients  $\beta_j$ . Le nombre entier  $q$  est ici la longueur du dernier retard; dans certains cas, imaginer que  $q$  est infini peut avoir du sens, mais nous supposons pour l'instant qu'il prend une valeur finie. La fonction de régression pourrait tout à fait dépendre d'autres variables explicatives, mais nous ignorons cette possibilité pour conserver une notation simple.

Le problème évident avec un modèle tel que (19.04) est que, parce que  $x_t$  sera souvent fortement corrélé à  $x_{t-1}$ ,  $x_{t-2}$ , et ainsi de suite, les estimations par moindres carrés des coefficients  $\beta_j$  tendront à être assez imprécises. De nombreux moyens pour manipuler ce problème furent proposés et nous en parlerons brièvement. La première chose à reconnaître est, malgré tout, que cela pourrait ne pas être un problème. Souvent, ce ne sont pas les coefficients individuels qui nous intéressent mais leur somme, disons  $\gamma$ , qui mesure l'effet de long terme sur  $y_t$  d'une variation donnée de  $x_t$ . Même lorsque les  $\beta_j$  individuels sont estimés de façon imprécise, leur somme peut être estimée avec suffisamment de précision.

Posons  $\mathbf{V}(\hat{\beta})$  la matrice de covariances du vecteur  $\hat{\beta}$  des estimations par moindres carrés dont l'élément type est  $\hat{\beta}_j$ . Alors, si  $\hat{\gamma}$  désigne la somme des  $\hat{\beta}_j$ , la variance de  $\hat{\gamma}$  est

$$V(\hat{\gamma}) = \mathbf{1}^\top \mathbf{V}(\hat{\beta}) \mathbf{1} = \sum_{j=0}^q V(\hat{\beta}_j) + 2 \sum_{j=0}^q \sum_{k=0}^{j-1} \text{Cov}(\hat{\beta}_j, \hat{\beta}_k). \quad (19.05)$$

Si  $x_{t-j}$  est corrélé positivement à  $x_{t-k}$  pour tout  $j \neq k$ , les termes de covariance dans (19.05) seront généralement négatifs. Lorsqu'ils sont importants et négatifs, comme c'est souvent le cas,  $V(\hat{\gamma})$  peut être plus petite que la somme des  $V(\hat{\beta}_j)$  ou même que chaque  $V(\hat{\beta}_j)$ .

Si c'est le paramètre  $\gamma$  qui nous intéresse plutôt que les  $\beta_j$  individuels, l'approche la plus simple consiste à estimer une version reparamétrisée de (19.04) par moindres carrés. La version reparamétrisée est

$$y_t = \alpha + \gamma x_t + \sum_{j=1}^q \beta_j (x_{t-j} - x_t) + u_t. \quad (19.06)$$

Il est aisé de vérifier que le coefficient  $\gamma$  associé à  $x_t$  dans (19.06) est en fait égal à la somme des  $\beta_j$  dans (19.04). L'avantage de cette reparamétrisation

est que l'écart type de  $\hat{\gamma}$  est immédiatement disponible dans les résultats de la régression.

Si notre intérêt se focalise sur les  $\beta_j$ , la colinéarité peut être un problème urgent. De nombreux moyens d'aborder ce problème furent proposés. Certains impliquent l'imposition de contraintes sur les paramètres de (19.04), alors que d'autres impliquent l'estimation de modèles pour lesquels une ou plusieurs retards de la variables dépendantes apparaissent dans l'ensemble des régresseurs. Cette dernière approche est fondamentalement différente de la première, et sera traitée dans la section qui suit. L'exemple le plus connu de la première approche consiste à employer ce que l'on nomme les **retards échelonnés polynomiaux**, ou **PDL**. Ces derniers sont quelquefois appelés **retards d'Almon** à la suite de l'article d'Almon (1965) à l'occasion duquel ils furent proposés pour la première fois.

Dans un polynôme de retards échelonnés, les coefficients  $\beta_j$  de (19.04) doivent se situer dans un polynôme de degré  $d$  donné. Ce polynôme peut éventuellement être soumis à des contraintes ultérieures, telles que les contraintes des portant sur les points terminaux. A titre d'exemple simple, si le polynôme était du second degré, sans contrainte ultérieure, nous aurions

$$\beta_j = \eta_0 + \eta_1 j + \eta_2 j^2 \quad \text{pour } j = 0, \dots, q. \quad (19.07)$$

A condition que  $q > 2$ , il y aura moins de paramètres  $\eta_i$  que  $\beta_j$ . Nous voyons par conséquent que (19.07) impose  $q - 2$  contraintes sur les  $\beta_j$ 's.

L'estimation d'un modèle soumis à des contraintes imposées par un PDL est conceptuellement assez immédiate. Par exemple, pour estimer (19.04) soumis à (19.07), nous remplacerions simplement les  $\beta_j$  par  $\eta_0 + \eta_1 j + \eta_2 j^2$ . Cela entraînerait

$$\begin{aligned} y_t &= \alpha + \eta_0 \sum_{j=0}^q x_{t-j} + \eta_1 \sum_{j=0}^q j x_{t-j} + \eta_2 \sum_{j=0}^q j^2 x_{t-j} + u_t \\ &= \alpha + \eta_0 z_{t0} + \eta_1 z_{t1} + \eta_2 z_{t2} + u_t. \end{aligned} \quad (19.08)$$

C'est simplement un modèle de régression linéaire avec trois nouveaux régresseurs  $z_{ti}$  qui sont des transformations des  $q + 1$  régresseurs d'origine, en plus de la constante. Ceci est un exemple de modèle  $\text{PDL}(q, 2)$ . Pour un modèle  $\text{PDL}(q, d)$ , qui doit toujours être tel que  $d < q$ , il y aurait  $d + 1$  régresseurs.

Les contraintes imposées aux  $\beta_j$  sont simplement des contraintes *linéaires*. La résolution de (19.07) nous montre que

$$\begin{aligned} -\beta_3 + 3\beta_2 - 3\beta_1 + \beta_0 &= 0, \\ -\beta_4 + 3\beta_3 - 3\beta_2 + \beta_1 &= 0, \\ -\beta_5 + 3\beta_4 - 3\beta_3 + \beta_2 &= 0, \text{ et ainsi de suite.} \end{aligned}$$

On peut écrire ces contraintes sous la forme  $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$ , où la matrice  $\mathbf{R}$  serait dans ce cas

$$\mathbf{R} = \begin{bmatrix} 1 & -3 & 3 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -3 & 3 & -1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -3 & 3 & -1 \end{bmatrix}.$$

Puisque les contraintes sont linéaires, on peut les tester facilement. On peut utiliser soit un test en  $F$  habituel, soit sa version robuste à l'hétéroscédasticité (voir Section 11.6). Le modèle contraint est (19.08), le modèle non contraint est (19.04), et le nombre de contraintes dans ce cas est  $q - 2$ . De façon plus générale, pour un modèle PDL( $q, d$ ), il y aura  $q - d$  contraintes.

Il faudrait *toujours* tester les contraintes imposées par n'importe quel type de PDL avant d'accepter, même à titre provisoire, un modèle qui incorpore ces contraintes. Ces contraintes sont de deux natures. Il y a la contrainte de la longueur du dernier retard qui ne doit pas être supérieure à  $q$ . Puis il y a les contraintes futures qui sont imposées par le PDL, quelles qu'elles soient. Pour une valeur de  $q$  donnée, la réduction du degré du polynôme de  $d$  à  $d - 1$  aboutit à un modèle plus restrictif. Cependant, pour un degré donné du polynôme, la réduction de  $q$  produit simplement un modèle différent, non emboîté, qui peut s'ajuster mieux ou plus mal aux données. Ainsi, on peut tester un modèle PDL( $q, d$ ) contre un modèle PDL( $q, d + 1$ ) en utilisant un test en  $F$  ordinaire, mais on ne peut pas tester un modèle PDL( $q, d$ ) contre un modèle PDL( $q + 1, d$ ) avec le même instrument. La meilleure approche consiste sans doute à se poser en premier le problème de la longueur du retard, en débutant par une valeur importante de  $q$  et en examinant la détérioration de la qualité de l'ajustement du modèle en diminuant sa valeur, sans imposer aucune contrainte sur la forme des retards échelonnés. Une fois que  $q$  est déterminé, on peut ensuite tenter de déterminer  $d$ , une fois encore en débutant avec une valeur importante et en la réduisant au fur et à mesure. Un excellent exemple empirique est donné par Sargan (1980c). La spécification d'un modèle final dans cette optique est un exemple de prétest dont nous avons discuté dans la Section 3.7; consulter Trivedi (1978).

La plupart des progiciels d'économétrie permettent aux utilisateurs de spécifier des modèles qui incluent des PDL et d'estimer de tels modèles avec des OLS, des IV, et quelquefois d'autres formes d'estimations. Ces mises en oeuvre sont de façon typique beaucoup plus sophistiquées que notre discussion n'a pu le suggérer jusqu'ici. Par exemple, elles permettent souvent à l'utilisateur de spécifier des contraintes additionnelles sur la forme des retards telles que les contraintes  $\beta_q = 0$ . Plus important encore, les bons progiciels utilisent des familles de polynômes plus sophistiquées que celles que nous avons décrites. Le problème avec ces dernières est que les variables  $z_{ti}$  tendent à être fortement corrélées entre elles. Cela peut provoquer une singularité numérique de la matrice  $\mathbf{X}^\top \mathbf{X}$ . Avec l'aide d'autres types de polynômes, tels que les

polynômes orthogonaux, on peut réduire en grande partie cette corrélation, et, par conséquent, éliminer ce genre de problème numérique. Les références à consulter sont Cooper (1972b), Trivedi et Pagan (1979), Sargan (1980c), et Pagano et Hartley (1981).

Shiller (1973) proposa une variante intéressante de l'approche PDL. Comme nous l'avons vu, les contraintes imposées par un PDL peuvent toujours s'écrire comme  $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$  pour une matrice  $\mathbf{R}$  de dimension  $r \times k$  convenablement définie. Ici  $r = q - d$  et  $k$  est le nombre d'éléments de  $\boldsymbol{\beta}$ , et il sera généralement supérieur à  $q + 1$  s'il y a des régresseurs en plus de la constante et des retards de  $x_t$ . Shiller suggéra que, au lieu de demander une vérification exacte de ces contraintes, nous requérions seulement qu'elles soient *approximatives*. Ainsi, au lieu de stipuler que chaque ligne de  $\mathbf{R}\boldsymbol{\beta}$  soit nulle, il proposa qu'elle soit égale à une variable aléatoire d'espérance nulle et de variance définie. L'un des avantages de cette approche est que  $d$  peut être très faible sans pour cela imposer des contraintes excessivement fortes sur les données. Puisque les estimations n'ont pas besoin de se conformer exactement à la forme du polynôme,  $d = 2$  est dans la plupart des cas une situation adéquate.

Ce genre de contrainte est appelé **contrainte stochastique**, parce qu'elle n'est pas sensée être vérifiée exactement. Les contraintes stochastiques sont très différentes de n'importe quel autre type de contraintes dont nous avons discuté. Dans de nombreuses situations, elles paraissent assez plausibles, à l'inverse des contraintes exactes, qui semblent être souvent excessivement fortes. Dans le cas du PDL, par exemple, il est sûrement peu probable que les  $\beta_j$  se situent réellement dans un polynôme de degré quelconque, mais il est assez probable de croire qu'ils se situent relativement près d'un tel polynôme. Il est aisé conceptuellement, mais plus difficile lors des phases de calcul, de traiter des contraintes stochastiques, ou n'importe quelle autre sorte d'information stochastique a priori, si l'on adopte un point de vue bayésien; voir Zellner (1971) et Drèze et Richard (1983). À l'inverse, il est facile de faire des calculs avec de telles contraintes, mais leur manipulation est conceptuellement plus délicate lorsque l'on reste dans une structure classique. C'est dans cette dernière que nous nous situerons, pour traiter les calculs, en évitant toute discussion relative aux difficultés conceptuelles.

La technique d'estimation suggérée par Shiller emploie un cas particulier de ce que Theil et Goldberger (1961) et Theil (1963) appellent une **estimation mixte**. L'estimation mixte est un moyen très simple de combiner des informations d'échantillon avec des informations stochastiques a priori. On peut imaginer que c'est une approximation d'une procédure qui a toutes les caractéristiques d'une estimation bayésienne. Le cas le plus simple pour lequel on justifie une estimation mixte est le cas dans lequel, avant d'entreprendre l'estimation d'un quelconque modèle, on aurait obtenu des estimations préalables d'un ou de plusieurs paramètres du modèle, par l'usage d'un ensemble d'informations totalement indépendantes. Pour faire simple,



supposons que le modèle qu'il s'agit d'estimer est le modèle de régression linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma_u^2 \mathbf{I}), \quad (19.09)$$

où  $\boldsymbol{\beta}$  est un vecteur à  $k$  composantes. Supposons ensuite qu'un vecteur d'estimations a priori  $\check{\boldsymbol{\beta}}$  soit disponible, avec sa *véritable* matrice de covariance  $\mathbf{V}(\check{\boldsymbol{\beta}})$ . On peut exprimer la relation entre ces estimations et le vecteur paramétrique inconnu  $\boldsymbol{\beta}$  comme

$$\check{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{v}, \quad E(\mathbf{v}\mathbf{v}^\top) = \mathbf{V}(\check{\boldsymbol{\beta}}) \equiv \boldsymbol{\eta}^{-1}(\boldsymbol{\eta}^\top)^{-1}. \quad (19.10)$$

Le membre de droite de l'expression pour la matrice de covariance fait usage d'un résultat standard sur les matrices définies positives, que nous avons vu dans le Chapitre 9. En prémultipliant chaque membre de (19.10) par la matrice  $\boldsymbol{\eta}$  de dimension  $k \times k$ , le résultat est

$$\boldsymbol{\eta}\check{\boldsymbol{\beta}} = \boldsymbol{\eta}\boldsymbol{\beta} + \mathbf{e}, \quad E(\mathbf{e}\mathbf{e}^\top) = \mathbf{I}. \quad (19.11)$$

Cela ressemble à une régression linéaire avec  $k$  observations et  $k$  variables indépendantes. La régressande est  $\boldsymbol{\eta}\check{\boldsymbol{\beta}}$ , et la matrice de covariance des aléas est  $\mathbf{I}$ .

Il devrait être aisé de voir comment on peut utiliser l'information contenue dans  $\check{\boldsymbol{\beta}}$  pour améliorer nos estimations de  $\boldsymbol{\beta}$ . Il suffit d'estimer une unique régression GLS à  $n + k$  observations, où  $n$  d'entre elles correspondent aux observations de notre échantillon et où  $k$  d'entre elles correspondent à (19.11). On peut écrire cette régression comme

$$\begin{bmatrix} \mathbf{y} \\ \sigma_u \boldsymbol{\eta}\check{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \sigma_u \boldsymbol{\eta} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u} \\ \sigma_u \mathbf{e} \end{bmatrix}. \quad (19.12)$$

Les aléas de cette régression sont i.i.d. et ont une variance égale à  $\sigma_u^2$ . La régression (19.12) suppose que nous connaissons  $\sigma_u$ , puisqu'il faut multiplier les  $k$  dernières observations par cette quantité de façon à garantir qu'elles ont le même poids relativement aux  $n$  premières observations. Asymptotiquement bien sûr, nous aurons les mêmes résultats si nous employons n'importe quelle estimation convergente de  $\sigma_u$ .

Dans cet exemple, l'estimation mixte ne prête pas trop à controverse. C'est simplement un moyen pratique de prendre en compte les estimations préalables lorsque l'on utilise un nouvel ensemble de données. Dans le cas des retards échelonnés, par contre, l'information a priori sur  $\boldsymbol{\beta}$  ne provient pas d'une estimation préalable. Au lieu de cela, c'est un ensemble de contraintes stochastiques, que Shiller appela une **information a priori de régularité** parce qu'il reflète la croyance qui veut que les coefficients  $\beta_j$  d'un retard échelonné devraient varier sans à-coups en fonction de  $j$ . Ces contraintes peuvent

paraître raisonnables au chercheur, mais elles ne se basent pas sur les données. Dans le cas général, on peut écrire les contraintes stochastiques comme

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{v}, \quad \mathbf{v} \sim N(\mathbf{0}, \sigma_v^2 \mathbf{I}). \quad (19.13)$$

Cette formulation autorise un éventail très large de contraintes linéaires sur  $\boldsymbol{\beta}$  comprend, en tant que cas particulier, l'imposition d'informations a priori de régularité sur les coefficients d'un retard échelonné. La matrice  $\mathbf{R}$  est de dimension  $r \times k$  et, dans le cas d'informations a priori de régularité, elle aura  $r = q - d$  lignes.

Pour pouvoir estimer (19.09) en imposant les contraintes stochastiques (19.13), nous réécrivons simplement ces dernières comme  $\mathbf{0} = \mathbf{R}\boldsymbol{\beta} + \mathbf{v}$ , comme nous l'avons fait dans (19.12). Les restrictions ressemblent alors aux observations d'une régression. Puis, nous empilons les véritables observations sur les observations artificielles. Cela donne

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{R} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}. \quad (19.14)$$

En fait, nous avons ajouté  $r$  observations supplémentaires à l'ensemble des données d'origine. La variance des "aléas" associés à ces observations supplémentaires est  $\sigma_v^2$ , alors que celle des aléas naturels est  $\sigma_u^2$ .

Posons maintenant  $\lambda \equiv \sigma_u / \sigma_v$ . Si  $\lambda$  était connu, l'estimation par GLS de (19.14) serait équivalente à l'estimation par OLS du modèle

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \lambda \mathbf{R} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u} \\ \lambda \mathbf{v} \end{bmatrix}. \quad (19.15)$$

L'estimation OLS de  $\boldsymbol{\beta}$  à partir de (19.15) est

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X} + \lambda^2 \mathbf{R}^\top \mathbf{R})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Il est facile de calculer cette expression, et il est aisé de la comprendre. Comme  $\sigma_v \rightarrow \infty$ ,  $\lambda \rightarrow 0$  et  $\tilde{\boldsymbol{\beta}} \rightarrow \hat{\boldsymbol{\beta}}$ . Ainsi, au fur et à mesure que la masse d'information contenue dans les restrictions stochastiques tend vers zéro, l'estimation mixte  $\tilde{\boldsymbol{\beta}}$  tend vers l'estimation OLS  $\hat{\boldsymbol{\beta}}$ . Dans le cas extrême opposé,  $\lambda \rightarrow \infty$  et  $\tilde{\boldsymbol{\beta}}$  converge vers un ensemble d'estimations qui satisfait les contraintes  $\mathbf{R}\boldsymbol{\beta} = \mathbf{0}$  au fur et à mesure que  $\sigma_v \rightarrow 0$ . Ce dernier résultat se comprend assez vite. Puisque  $r < k$ , il est toujours possible d'ajuster les  $r$  dernière lignes de (19.15) à la perfection en choisissant  $\tilde{\boldsymbol{\beta}}$  pour satisfaire les contraintes avec exactitude. Comme  $\lambda \rightarrow \infty$ , la SSR pour (19.15) s'accroîtra infiniment si les  $r$  dernières lignes ne s'ajustent pas parfaitement. Ainsi, comme on peut le voir à l'aide de l'algèbre matriciel fastidieuse, la limite de  $\tilde{\boldsymbol{\beta}}$  lorsque  $\lambda \rightarrow \infty$  est précisément l'estimateur des moindres carrés qui provient de l'imposition exacte des contraintes.

Le problème majeur de cette procédure est que  $\lambda$  ne sera jamais connu. Même si l'on désire spécifier  $\sigma_v$  a priori, ce qui peut ne pas être simple à faire,  $\sigma_u$  devra tout de même être estimée. Il existe des moyens variés de traiter ce problème — voir Shiller (1973) et Taylor (1974) — mais aucun d'entre eux n'est entièrement satisfaisant. Pour l'essentiel, il s'agit d'estimer  $\sigma_u$  à partir de l'estimation non contrainte de (19.09), soit en prenant une valeur pour  $\sigma_v$  soit en estimant  $\sigma_v$  à partir des estimations non contraintes de  $\beta$ , et de construire une estimation de  $\lambda$ . Cela transforme la procédure d'estimation mixte en une forme d'estimation par GLS faisables. Asymptotiquement, cela produira les mêmes estimations que si  $\lambda$  était connu, mais ses performances avec des échantillons finis peuvent ne pas être aussi bonnes.

Il faudrait toujours tester des contraintes stochastiques avant d'accepter des estimations basées sur ces contraintes. Puisque l'imposition de telles restrictions est équivalente à l'addition d'observations factices, le moyen évident de les tester est d'utiliser un test standard pour l'égalité de deux ensembles de paramètres de régression (Section 11.2). On peut voir (19.15) comme un modèle pour l'échantillon entier (augmenté), où  $\beta$  est contraint à être identique pour les  $n$  premières observations et les  $r$  observations restantes. L'estimation de (19.15) produit la somme des résidus au carré contrainte RSSR nécessaire à la construction d'un test en  $F$ . Puisque  $r < k$ , toute tentative d'estimation des paramètres utilisant le second sous-échantillon uniquement entraînera des estimations qui s'ajustent parfaitement. Ainsi la somme des résidus au carré non contrainte USSR nécessaire à la construction d'un  $F$  de Fisher est simplement la somme des résidus au carré de l'estimation par OLS de (19.09). Le nombre de degrés de liberté pour le test est  $r$ , et par conséquent le  $F$  de Fisher est simplement

$$\frac{(\text{RSSR} - \text{USSR})/r}{\text{USSR}/(n - k)}.$$

Bien évidemment, on pourrait utiliser une quelconque autre forme de statistique de test, telle que celle basée sur la HRGNR (11.66), au lieu du  $F$  de Fisher. Si le test rejette l'hypothèse nulle de constance de  $\beta$  sur l'échantillon des observations et sur les observations factices, il faudrait soit accroître la valeur de  $\sigma_v$  soit changer la forme de la matrice  $\mathbf{R}$ , probablement en augmentant  $d$ .

Bien que les retards échelonnés polynomiaux, qu'ils soient imposés en tant que contraintes exactes ou en tant que contraintes stochastiques, puissent être utiles lorsqu'un modèle tel que (19.04) est inadapté, ce ne sont pas des modélisations toujours bien appropriées. Le problème est que (19.04) n'est pas un modèle *dynamique*. Bien que  $y_t$  dépende de valeurs retardées de  $x_t$ , elle ne dépend pas de ses propres valeurs retardées. Par conséquent, seule la valeur courante de  $u_t$  affecte  $y_t$ . Mais si l'on pense que l'aléa doit représenter l'influence combinée de nombreuses variables dont on ne peut empêcher l'omission de la régression, cela devrait paraître étrange. Après

tout, si  $x_t$  affecte  $y_t$  au travers d'un retard échelonné, comment justifier que les variables reléguées dans l'aléa n'en fassent pas de même? Cet argument suggère que les aléas dans un modèle comparable à (19.04) peuvent être très souvent corrélés en série. Bien sûr, on peut modéliser les  $u_t$  pour les faire obéir à un quelconque processus ARMA. Mais la meilleure approche consistera souvent à reformuler le modèle originel. Nous allons voir comment dans la prochaine section.

## 19.4 MODÈLES DE RÉGRESSION DYNAMIQUES

Tout modèle de régression dans lequel la fonction de régression dépend des valeurs retardées d'une ou de plusieurs variables dépendantes est appelé **modèle dynamique**. Les seuls modèles dynamiques dont nous ayons discuté jusqu'à présent sont les modèles à erreurs corrélées en série (Chapitre 10); après transformation, les modèles à erreurs AR ou MA impliquent des retards de la variable dépendante. Ces modèles peuvent paraître artificiels, mais les modèles dynamiques peuvent survenir pour de nombreuses autres raisons.

Un modèle dynamique simple et très fréquent est le **modèle d'ajustement partiel**, dont l'histoire en économie remonte assez loin puisqu'il date de Nerlove (1958). Supposons que le niveau *désiré* d'une variable économique  $y_t$  quelconque soit  $y_t^*$ , qui est supposé être relié à un vecteur de variables explicatives exogènes  $\mathbf{X}_t$  comme suit:

$$y_t^* = \mathbf{X}_t \boldsymbol{\beta}^* + e_t. \quad (19.16)$$

A cause de certains coûts d'ajustement, les agents ne peuvent pas atteindre  $y_t^*$  à chaque période. Au lieu de cela,  $y_t$  s'ajuste, par hypothèse, vers  $y_t^*$  suivant l'équation

$$y_t - y_{t-1} = (1 - \delta)(y_t^* - y_{t-1}) + v_t. \quad (19.17)$$

La résolution de (19.16) et de (19.17) pour  $y_t$  nous permet d'obtenir

$$\begin{aligned} y_t &= y_{t-1} - (1 - \delta)y_{t-1} + (1 - \delta)\mathbf{X}_t \boldsymbol{\beta}^* + (1 - \delta)e_t + v_t \\ &= \mathbf{X}_t \boldsymbol{\beta} + \delta y_{t-1} + u_t, \end{aligned} \quad (19.18)$$

où  $\boldsymbol{\beta} \equiv (1 - \delta)\boldsymbol{\beta}^*$  et  $u_t \equiv (1 - \delta)e_t + v_t$ . Si l'on désire estimer  $\boldsymbol{\beta}^*$ , on peut aisément le faire à partir des estimations OLS de  $\boldsymbol{\beta}$  et  $\delta$ .

L'ajustement partiel n'est pertinent que si  $0 < \delta < 1$  et si, de plus,  $\delta$  n'est pas trop proche de 1, puisque dans le cas contraire la vitesse d'ajustement que la valeur du paramètre implique devient trop faible. On peut résoudre l'équation (19.18) pour  $y_t$  comme une fonction des valeurs courantes et passées de  $\mathbf{X}_t$  et  $u_t$ . Le résultat est

$$y_t = \sum_{j=0}^{\infty} \delta^j (\mathbf{X}_{t-j} \boldsymbol{\beta} + u_{t-j}). \quad (19.19)$$

Ainsi ce modèle corrige une défaillance majeure que nous avons déjà remarquée dans les modèles à retards échelonnés:  $y_t$  dépend maintenant autant des valeurs retardées de l'aléa  $u_t$  que des valeurs retardées des variables exogènes  $\mathbf{X}_t$ . Notons que la solution de (19.19) repose sur l'hypothèse que  $|\delta| < 1$ , qui est une condition de stationnarité pur ce modèle.

Le modèle d'ajustement partiel n'est qu'un des nombreux modèles économiques que l'on peut utiliser pour justifier la prise en compte d'un ou de plusieurs retards des variables dépendantes dans la fonction de régression. Dhrymes (1971) et Hendry, Pagan, et Sargan (1984) discutent de nombreux autres modèles. Nous n'essaierons pas de discuter de ces derniers. Par contre, nous nous concentrerons sur quelques résultats d'ordre général qui peuvent survenir lorsque l'on tente de spécifier et d'estimer des modèles de régression dynamiques.

Un problème qui se manifeste chaque fois que la matrice  $\mathbf{X}$  contient des variables dépendantes retardées est que les OLS ne produisent pas des estimations sans biais. Ce problème survient parce que  $\mathbf{X}$  est une matrice stochastique, dont certains éléments sont corrélés à quelques éléments de  $\mathbf{u}$ . Ainsi

$$E((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}) \neq (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{u}).$$

Le meilleur moyen d'apercevoir ce problème est de considérer un exemple très simple. Supposons que

$$y_t = \beta y_{t-1} + u_t, \quad |\beta| < 1, \quad u_t \sim \text{IID}(0, \sigma^2). \quad (19.20)$$

L'estimation OLS de  $\beta$  est

$$\hat{\beta} = \frac{\sum_{t=2}^n y_t y_{t-1}}{\sum_{t=2}^n y_{t-1}^2}. \quad (19.21)$$

Si l'on substitue (19.20) au numérateur de (19.21), on obtient

$$\hat{\beta} = \frac{\beta \sum_{t=2}^n y_{t-1}^2 + \sum_{t=2}^n u_t y_{t-1}}{\sum_{t=2}^n y_{t-1}^2} = \beta + \frac{\sum_{t=2}^n u_t y_{t-1}}{\sum_{t=2}^n y_{t-1}^2}. \quad (19.22)$$

Le second terme dans l'expression la plus à droite de (19.22) *n'est pas* d'espérance nulle, parce que le numérateur et le dénominateur ne sont pas indépendants. Son espérance est assez difficile à déterminer. Nous concluons que dans ce modèle, et dans tous les modèles pour lesquels il y a des variables dépendantes retardées, l'estimateur OLS est biaisé.

Evidemment, l'estimateur OLS  $\hat{\beta}$  est convergent comme des résultats établis antérieurement l'ont montré (Section 5.3). Si l'on divise à la fois le numérateur et le dénominateur du terme aléatoire du membre le plus à droite de (19.22) par  $n$  et si l'on prend le limites en probabilité, on obtient

$$\text{plim}_{n \rightarrow \infty} \hat{\beta} = \beta + \frac{\text{plim}_{n \rightarrow \infty} (n^{-1} \sum_{t=2}^n u_t y_{t-1})}{\text{plim}_{n \rightarrow \infty} (n^{-1} \sum_{t=2}^n y_{t-1}^2)} = \beta.$$

La limite en probabilité du numérateur est zéro. Cela provient du fait que  $E(u_t y_{t-1}) = 0$ , ce qui implique que  $n^{-1} \sum_{t=2}^n u_t y_{t-1}$  est simplement la moyenne de  $n$  quantités qui sont toutes d'espérance nulle, et que ces quantités sont de variance finie, ce qui est le cas puisque le fait que  $|\beta| < 1$  implique que le processus générateur des  $y_t$  est stationnaire. La limite en probabilité du numérateur est finie, ce qui nécessite à nouveau la stationnarité, et par conséquent le rapport des deux limites en probabilité est nul.

Même pour un modèle aussi simple que (19.20), les propriétés avec des échantillons finis de l'estimateur OLS  $\hat{\beta}$  sont assez difficiles à établir de façon analytique et elles dépendent de la valeur (inconnue) de  $\beta$ ; nous présenterons quelques résultats Monte Carlo dans le Chapitre 21. Dans des modèles plus compliqués, les chercheurs disposent de choix restreints et sont contraints de se rapporter à la théorie asymptotique. Cela n'est pas un mal en général, bien qu'il y ait un risque évident que des inférences non correctes soient produites, en particulier lorsque la taille de l'échantillon est faible ou que le modèle est presque non stationnaire.

Nous considérons maintenant une classe très étendue de modèles de régression linéaire dynamiques qui peuvent être très utiles dans la pratique. Ces modèles ne possèdent qu'une seule variable dépendante  $y_t$  et, pour simplifier la notation, une seule variable indépendante  $x_t$ . Un modèle **autorégressif à retards échelonnés**, ou modèle **ADL**, peut s'écrire comme

$$y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \sum_{j=0}^q \gamma_j x_{t-j} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2) \quad (19.23)$$

ou, en utilisant les opérateurs retard

$$A(L, \beta) y_t = \alpha + B(L, \gamma) x_t + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

Ici  $A(L, \beta)$  et  $B(L, \gamma)$  désignent les polynômes des opérateurs retards avec les coefficients respectifs  $\beta$  et  $\gamma$ . Parce qu'il y a  $p$  retards sur  $y_t$  et  $q$  retards sur  $x_t$ , on appelle quelquefois ces modèles les modèles **ADL(p, q)**. S'il y a des variables dépendantes additionnelles, ce qui sera en réalité le cas le plus fréquent, elles apparaîtront en tant que régresseurs additionnels dans (19.23).

Un cas particulièrement simple de (19.23), mais largement répandu, est le modèle **ADL(1, 1)**

$$y_t = \alpha + \beta_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + u_t. \quad (19.24)$$

Parce que la plupart des résultats qui sont vrais pour le modèle **ADL(1, 1)** sont également vrais, compte tenu de certaines modifications évidentes, pour le modèle plus général **ADL(p, q)**, nous bornerons notre discussion au cas particulier la plupart du temps.

de nombreux modèles pour séries temporelles que l'on rencontre couramment sont des cas spéciaux du modèle **ADL(1, 1)**. Un modèle de régression

statique est un cas particulier avec  $\beta_1 = \gamma_1 = 0$ , un modèle AR(1) univarié est un cas particulier avec  $\gamma_0 = \gamma_1 = 0$ , un modèle d'ajustement partiel est un cas particulier avec  $\gamma_1 = 0$ , un modèle statique à aléas AR(1) est un cas particulier avec  $\gamma_1 = -\beta_1\gamma_0$ , un modèle en différences premières est un cas particulier avec  $\beta_1 = 1$  et  $\gamma_1 = -\gamma_0$ , et ainsi de suite. Le modèle ADL(1, 1) fournit une alternative naturelle contre laquelle on peut tester n'importe lequel de ces cas particuliers. Un test des contraintes du facteur commun découlant des aléas obéissant à un processus AR(1) en est un exemple; voir la Section 10.9.

Examinons à présent comment  $x_t$  affecte  $y_t$  en longue période dans un modèle ADL(1, 1). Sans aléas,  $x_t$  et  $y_t$  convergeraient vers des valeurs de long terme stable  $x^*$  et  $y^*$  données par

$$y^* = \alpha + \beta_1 y^* + \gamma_0 x^* + \gamma_1 x^*.$$

En résolvant cette équation pour  $y^*$  en fonction de  $x^*$  on obtient

$$y^* = \frac{\alpha}{1 - \beta_1} + \frac{\gamma_0 + \gamma_1}{1 - \beta_1} x^* = \frac{\alpha}{1 - \beta_1} + \lambda x^*.$$

Nous voyons donc que la dérivée de  $y^*$  par rapport à  $x^*$  en longueur période (cette valeur correspondra à une élasticité si les deux séries sont exprimées en logarithmes) est

$$\lambda \equiv \frac{\gamma_0 + \gamma_1}{1 - \beta_1}. \quad (19.25)$$

A l'évidence, ce résultat est pertinent uniquement si  $|\beta_1| < 1$ , ce qui, comme on pourrait s'y attendre, est une condition de stabilité pour ce modèle.

L'une de ses caractéristiques intéressante et importante des modèles ADL est que l'on peut les écrire de différentes façons sans amoindrir leur faculté d'explication des données ou modifier les estimations par moindres carrés des coefficients auxquels on porte un intérêt. Par exemple, (19.24) peut être écrit selon toutes les formes qui suivent:

$$\Delta y_t = \alpha + (\beta_1 - 1)y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + u_t; \quad (19.26)$$

$$\Delta y_t = \alpha + (\beta_1 - 1)y_{t-1} + \gamma_0 \Delta x_t + (\gamma_0 + \gamma_1)x_{t-1} + u_t; \quad (19.27)$$

$$\Delta y_t = \alpha + (\beta_1 - 1)y_{t-1} - \gamma_1 \Delta x_t + (\gamma_0 + \gamma_1)x_t + u_t; \quad (19.28)$$

$$\begin{aligned} \Delta y_t &= \alpha + (\beta_1 - 1)(y_{t-1} - x_{t-1}) + \gamma_0 \Delta x_t \\ &\quad + (\gamma_0 + \gamma_1 + \beta_1 - 1)x_{t-1} + u_t; \end{aligned} \quad (19.29)$$

$$\Delta y_t = \alpha + (\beta_1 - 1)(y_{t-1} - \lambda x_{t-1}) + \gamma_0 \Delta x_t + u_t. \quad (19.30)$$

Ici  $\Delta$  est l'**opérateur des différences premières**:  $\Delta y_t \equiv y_t - y_{t-1}$ . Dans (19.30),  $\lambda$  est la paramètre défini dans (19.25). Le fait que (19.24) puisse être écrit sous différentes formes sans changer les estimations par moindres carrés est souvent très pratique. Par exemple, si l'on s'intéresse à la somme des  $\gamma_i$ , les

estimations et les écarts types s'obtiennent directement à partir de l'estimation par OLS de (19.27) ou (19.28), et si l'on porte un intérêt à  $\lambda$ , elles peuvent être obtenues par une estimation NLS de (19.30).

La plus intéressante des spécifications équivalentes (19.24) et (19.26)–(19.30) est sans doute (19.30), dans laquelle le modèle est écrit sous la forme que l'on appelle **forme à correction d'erreur**. Le paramètre  $\lambda$  apparaît directement dans cette forme du modèle. Bien que la forme à correction d'erreur soit non linéaire, l'estimation est malgré tout aisée parce que le modèle est simplement un modèle linéaire soumis à une contrainte non linéaire. La différence entre  $y_{t-1}$  et  $\lambda x_{t-1}$  mesure l'importance de la défaillance de la relation d'équilibre de long terme entre  $x_t$  et  $y_t$ . À ce titre,  $\beta_1 - 1$  est pour l'essentiel la même chose que le paramètre  $\delta - 1$  dans le modèle d'ajustement partiel. On appelle souvent le terme  $(\beta_1 - 1)(y_{t-1} - \lambda x_{t-1})$  qui apparaît dans (19.30) **terme de correction d'erreur**, et un modèle tel que (19.30) est parfois appelé **modèle à correction d'erreur**, ou **ECM**. Ces modèles furent utilisés pour la première fois par Hendry et Anderson (1977) et Davidson, Hendry, Srba, et Yeo (1978). Nous en discuterons en détail dans le prochain chapitre. Remarquons que le terme d'erreur est implicitement présent dans les autres versions de (19.24), puisque son coefficient associé peut être retrouvé à partir de celles-ci. Certains auteurs imposent la contrainte  $\lambda = 1$ , qui peut s'avérer raisonnable si  $x_t$  et  $y_t$  sont d'amplitudes comparables. Cela est équivalent à la contrainte  $\beta_1 + \gamma_0 + \gamma_1 = 1$  et peut donc être testé de façon assez simple par l'utilisation des  $t$  de Student ordinaires pour  $x_{t-1}$  dans (19.29).

Le point clef à retenir lorsque l'on tente de spécifier des modèles de régression dynamiques est qu'il existe en général un grand nombre de manières a priori plausibles de le faire. C'est une erreur grave que de limiter ses efforts sur un type particulier de modèles, tel que les modèles à retards échelonnés ou les modèles d'ajustement partiel. Parce qu'elle comporte tellement d'autres cas particuliers, la famille des modèles  $ADL(p, q)$  fournira souvent une bonne base de départ. Dans de nombreux cas, la spécification  $p = q = 1$  sera généralement suffisante, mais avec des données trimestrielles il serait sage de débiter avec  $p = q = 4$ . Dans le but d'obtenir un modèle raisonnablement économe et directement interprétable, il sera généralement nécessaire d'imposer un certain nombre de contraintes sur la spécification  $ADL(p, q)$  d'origine. Parce que les modèles  $ADL$  peuvent s'écrire de plusieurs manières différentes —souvenons-nous des modèles (19.24) et (19.26) à (19.30)— il y a également de nombreuses contraintes différentes que l'on pourrait imposer.

Notre discussion sur les modèles de régression dynamiques dut assez rapide. Pour des traitements plus pointus, consulter Hendry, Pagan, et Sargan (1984) ou Banerjee, Dolado, Galbraith, et Hendry (1993).



## 19.5 AUTORÉGRESSIONS VECTORIELLES

Dans le Chapitre 10, nous avons introduit les modèles AR, MA et ARMA pour des séries temporelles univariées. Comme on pourrait s'y attendre, il existe des versions multivariées de tous ces modèles. Nous ne tenterons pas de discuter des modèles à moyenne mobile vectoriels ou des modèles ARMA vectoriels, parce que ceux-ci peuvent être relativement compliqués à traiter; consulter Fuller (1976) ou Harvey (1981, 1989). Toutefois, dans cette section, nous verrons brièvement les **modèles autorégressifs vectoriels**, que l'on connaît également sous le nom d'**autorégressions vectorielles** ou **VAR**. Ceux-ci représentent le genre le plus simple de modèle de séries temporelles multivariées à estimer, et ils ont été largement employés en économie ces dernières années.

Supposons que le vecteur ligne  $\mathbf{Y}_t$  de dimension  $1 \times m$  désigne la  $t^{\text{ième}}$  observation d'un ensemble de variables. Alors un modèle autorégressif vectoriel d'ordre  $p$ , ou **VAR**( $p$ ) pour faire court, peut s'écrire comme

$$\mathbf{Y}_t = \boldsymbol{\alpha} + \mathbf{Y}_{t-1}\boldsymbol{\Phi}_1 + \cdots + \mathbf{Y}_{t-p}\boldsymbol{\Phi}_p + \mathbf{U}_t, \quad \mathbf{U}_t \sim \text{IID}(\mathbf{0}, \boldsymbol{\Omega}), \quad (19.31)$$

où  $\boldsymbol{\alpha}$  est un vecteur ligne à  $m$  composantes, et  $\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2$  jusqu'à  $\boldsymbol{\Phi}_p$  sont des matrices de dimension  $m \times m$  des coefficients qu'il faut estimer. Si  $y_{ti}$  désigne le  $i^{\text{ième}}$  élément de  $\mathbf{Y}_t$  et si  $\phi_{j,ki}$  désigne le  $ki^{\text{ième}}$  élément de  $\boldsymbol{\Phi}_j$ , la colonne  $i$  de (19.31) peut s'écrire comme

$$y_{ti} = \alpha_i + \sum_{j=1}^p \sum_{k=1}^m y_{t-j,k} \phi_{j,ki} + u_{ti}. \quad (19.32)$$

C'est simplement une régression linéaire, dans laquelle  $y_{ti}$  dépend d'une constante et des retards 1 à  $p$  des  $m$  variables du système. Ainsi (19.31) prend la forme d'un système SUR (Section 9.8).

Parce qu'exactement les mêmes variables apparaissent dans le membre de droite de (19.32) quel que soit  $i$ , les estimations OLS pour chaque équation sont identiques aux estimations GLS pour (19.31) prises ensembles. Cela est une conséquence du Théorème de Kruskal, ainsi que nous l'avons démontré à la Section 9.8. Ainsi il est très aisé d'estimer une VAR: on applique simplement les OLS à chaque équation de façon isolée. L'estimation est très rapide si le logiciel utilise le fait que chaque équation implique exactement le même ensemble de régresseurs.

L'usage des modèles VAR fut préconisé, notamment par Sims (1980), comme un moyen d'estimer des relations dynamiques entre des variables endogènes jointes sans avoir à imposer de fortes contraintes préalables. Des articles empiriques fondés sur cette approche furent écrits par Litterman et Weiss (1985) et Reagan et Sheehan (1985). L'avantage principal de cette approche est que le chercheur n'a pas besoin de décider quelles sont les variables

endogènes. De plus, tous les problèmes associés aux modèles d'équations simultanées sont contournés parce que les VAR ne contiennent aucune variable courante parmi les régresseurs. D'un autre côté, les VAR tendent à nécessiter l'estimation d'un grand nombre de paramètres,  $m + pm^2$  pour être précis, et, par conséquent, chaque paramètre individuel a tendance à être souvent estimé de façon assez imprécise. Nous reviendrons sur ce point plus tard.

Bien que le modèle VAR ne contienne pas de variables courante parmi les régresseurs, les corrélations contemporaines sont prises en compte de façon implicite par la matrice  $\Omega$ . Cette matrice est intéressante à plusieurs titres, et pas des moindres parce que, si les aléas sont supposés être normalement distribués, la fonction de log-vraisemblance pour le modèle VAR( $p$ ) (19.31), concentrée par rapport à  $\Omega$ , est simplement

$$\ell(\mathbf{Y}, \alpha, \Phi_1 \cdots \Phi_p) = C - \frac{n}{2} \log |\Omega(\alpha, \Phi_1 \cdots \Phi_p)|.$$

Ici  $\Omega(\alpha, \Phi_1 \cdots \Phi_p)$  signifie que l'on prend la valeur de  $\Omega$  qui maximise la log-vraisemblance conditionnellement à  $\alpha$  et aux  $\Phi_i$ , et  $\mathbf{Y}$  représente la matrice dont la ligne type est  $\mathbf{Y}_t$ . Ce résultat est une application des résultats relatifs aux fonctions de log-vraisemblance concentrées pour les modèles multivariés que nous avons dérivés à la Section 9.9;

Il est aisé de voir que  $\Omega(\alpha, \Phi_1 \cdots \Phi_p)$  est égale à

$$\frac{1}{n} \sum_{t=1}^n (\mathbf{Y}_t - \alpha - \mathbf{Y}_{t-1}\Phi_1 \cdots - \mathbf{Y}_{t-p}\Phi_p)^\top (\mathbf{Y}_t - \alpha - \mathbf{Y}_{t-1}\Phi_1 \cdots - \mathbf{Y}_{t-p}\Phi_p),$$

où nous avons supposé implicitement que les  $p$  observations antérieures à celles de l'échantillon sont disponibles, ce qui implique que les  $n$  observations soient employées pour l'estimation. Si  $\hat{\mathbf{U}}_t$  désigne le vecteur ligne à  $m$  éléments des résidus OLS pour l'observation  $t$ , alors

$$\Omega(\hat{\alpha}, \hat{\Phi}_1 \cdots \hat{\Phi}_p) \equiv \hat{\Omega} = \frac{1}{n} \sum_{t=1}^n \hat{\mathbf{U}}_t^\top \hat{\mathbf{U}}_t.$$

Par conséquent la valeur maximisée de la fonction de log-vraisemblance est

$$\ell(\mathbf{Y}, \hat{\alpha}, \hat{\Phi}_1 \cdots \hat{\Phi}_p) = C - \frac{n}{2} \log |\hat{\Omega}|.$$

Lorsque nous spécifions une modélisation VAR, il est important de déterminer la longueur des retards qu'il est nécessaire d'inclure. Si l'on désire tester l'hypothèse nulle que le retard le plus long dans le système est  $p$  contre l'hypothèse alternative que c'est  $p + 1$ , le moyen le plus facile de procéder est probablement de calculer la statistique LR

$$n(\log |\hat{\Omega}(p)| - \log |\hat{\Omega}(p+1)|),$$

avec une notation qui est très explicite. La distribution asymptotique de cette statistique de test sera le  $\chi^2(m^2)$ . Cependant, à moins que la taille  $n$  de l'échantillon ne soit très grande par rapport au nombre des paramètres dans le système ( $m + pm^2$  sous l'hypothèse nulle,  $m + (p + 1)m^2$  sous l'hypothèse alternative) la distribution avec des échantillons finis de cette statistique de test peut différer substantiellement de sa distribution asymptotique.

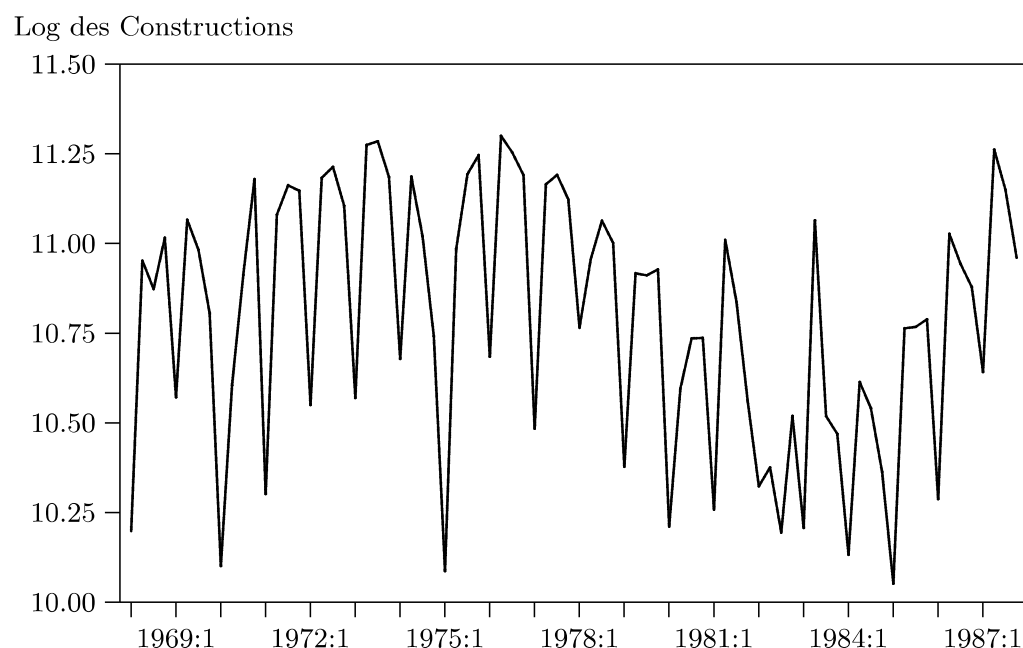
L'un des usages des modèles VAR est le test de l'hypothèse nulle qu'une quelconque variable ne possède pas de causalité au sens de Granger sur une autre variable. Nous avons discuté du concept de causalité au sens de Granger dans la Section 18. Dans le contexte d'une VAR, on dit qu'il y a causalité au sens de Granger entre  $y_{t1}$  et  $y_{t2}$  si les valeurs retardées de  $y_{t1}$  sont significatives dans l'équation de  $y_{t2}$ . D'un autre côté, l'hypothèse nulle que  $y_{t1}$  ne cause pas  $y_{t2}$  au sens de Granger ne peut pas être rejetée si toutes les valeurs retardées de  $y_{t1}$  sont conjointement sans pertinence dans l'équation de  $y_{t2}$ . Ainsi on peut facilement tester l'hypothèse nulle que n'importe quelle variables dans une VAR( $p$ ) n'a pas de causalité au sens de Granger sur n'importe quelle autre variable en exécutant un test en  $F$  asymptotique avec  $p$  et  $n - (1 + pm)$  degrés de liberté.<sup>2</sup> A l'évidence, tous les résultats dépendent de l'hypothèse maintenue que toutes les variables pertinentes ont été incluses dans la VAR. Si une variable  $y_{t3}$  était omise de la VAR, nous concluerions à tort que  $y_{t1}$  cause  $y_{t2}$  au sens de Granger, alors qu'en réalité  $y_{t1}$  n'explique pas du tout  $y_{t2}$  indépendamment de son effet à travers la variable omise.

Comme nous le remarquons déjà, un problème pratique particulièrement délicat avec les VAR est qu'elles réclament généralement l'estimation d'un nombre de paramètre qui est important relativement à la taille de l'échantillon. Litterman (1979, 1986) suggéra que si l'objectif est l'utilisation d'une VAR pour la prévision, on peut résoudre ce problème en imposant des contraintes aléatoires, très similaires à celles que nous avons vues dans la Section 19.2 et dont le but était d'imposer des informations a priori de régularité sur les retards échelonnés. Par exemple, on pourrait imposer l'information a priori que tous les coefficients sont d'espérance nulle et de variance assez forte, excepté pour le coefficient associé à  $y_{t-1,i}$  dans l'équation pour  $y_{ti}$ . Litterman proposa une procédure d'estimation mixte similaire à celle dont nous avons discuté lors de la Section 19.2, et rapporta que ces VAR "bayésiennes" produisaient de meilleures prévisions que les VAR non contraintes conventionnelles.

## 19.6 L'AJUSTEMENT SAISONNIER

De nombreuses séries temporelles économiques tendent à suivre un modèle régulier à travers le déroulement de chaque année. On appelle ce genre de comportement une **variation saisonnière** ou **saisonnalité**. Il peut provenir de

<sup>2</sup> Les propriétés des différents tests de causalité, incluant celui-ci, furent étudiées par Geweke, Meese, et Dent (1983).



**Figure 19.1** Constructions de bâtiments au Canada, 1968–1987

conditions climatiques saisonnières régulières ou d'habitudes sociales telles que les jours fériés légaux, les vacances en été et d'autres. La présence de saisonnalité a des implications importantes dans les travaux économétriques appliqués qui utilisent des données chronologiques. Au mieux, lorsque nous parvenons à modéliser la saisonnalité de manière explicite, cela complique le travail dans une large mesure. Au pire, l'utilisation de données corrigées des variations saisonnières de façon mécanique peut réduire drastiquement notre capacité à pratiquer des inférences correctes sur des relations économiques.

Pour clarifier les idées, considérons la Figure 19.1, qui présente le logarithme des constructions de bâtiments au Canada, en données trimestrielles, pour la période 1968:1 à 1987:4.<sup>3</sup> Il est clair que la variation saisonnière dans cette série est très prononcée. Les constructions de bâtiments tendent à être plus faibles lors du premier trimestre que lors des autres, sans doute parce que les conditions climatiques de l'hiver rendent les travaux difficiles en cette période de l'année. Malgré cela, le modèle de saisonnière paraît varier considérablement d'une année à l'autre, d'une manière que ne semble pas indépendante du niveau général des constructions d'immeubles. Dans l'année de récession de 1982, par exemple, il y a beaucoup moins de variations saisonnières que d'habitude, et le niveau le plus faible des constructions est enregistré pour le troisième trimestre au lieu du premier.

<sup>3</sup> Ces données sont issues de la base de données CANSIM des Statistiques Canadiennes. Elles correspondent aux logarithmes de la série numéro D2717.

Il existe deux visions assez divergentes sur la nature de la saisonnalité dans les données économiques. la première est que la variation saisonnière est une partie fondamentale de nombreuses séries économiques et, lorsqu'elle se manifeste, il faudrait essayer de l'expliquer. Ainsi, dans un monde idéal, un modèle économétrique pour une variable dépendante  $y_t$  devrait expliquer n'importe quelle variation saisonnière des variables indépendantes, sans doute en incluant des variables saisonnières muettes parmi elles. Hélas, comme nous allons le voir dans la section qui suit, cela rend la spécification et l'estimation économétrique des modèles pour séries mensuelles ou trimestrielles relativement compliquées.

La seconde interprétation, associée à Sims (1974), est que la saisonnalité est simplement un type de perturbation qui contamine les données économiques. La théorie économique n'est pas supposée expliquer ce bruit, qui, dans le cas de variables indépendantes, équivaut à un problème d'erreur dans les variables. On devrait par conséquent utiliser ce que l'on appelle les données **ajustées par saison**, c'est-à-dire des données qui ont été conditionnées d'une certaine façon de sorte qu'elle représentent ce que nous supposons que la série serait en l'absence de saisonnalité. En réalité, de nombreux bureaux d'études, en particulier aux Etats Unis, produisent uniquement des chiffres ajustés par saison pour de nombreuses séries. Dans cette section, nous allons discuter de la nature des procédures d'ajustement saisonnier et des conséquences de l'utilisation des données ajustées par saison.

L'idée d'ajuster par saison une série temporelle afin d'éliminer les effets de la saisonnalité est intuitivement attrayante mais assez difficile à rendre rigoureuse sans avoir à s'appuyer sur des hypothèses beaucoup trop irréalistes. L'ajustement saisonnier d'une série  $y_t$  est pertinent pour tout  $t$  on peut écrire  $y_t = y_t^* + y_t^s$ , où  $y_t^*$  est une série temporelle qui ne contient aucune variation saisonnière, et  $y_t^s$  est une série temporelle qui ne contient que des composantes saisonnières. Mais cela est une hypothèse extrême. Même si elle est vérifiée, il n'est pas nécessairement aisé de séparer  $y_t$  en  $y_t^*$  et  $y_t^s$ , ce qui est ce que les procédures d'ajustement saisonnier tentent d'accomplir.

Une approche de l'ajustement saisonnier, qui est très populaire parmi les économètres mais qui n'est presque jamais utilisée par les bureaux d'études statistiques, consiste à utiliser une régression par moindres carrés. Supposons pour être concret, que les données sont trimestrielles, et considérons les variables saisonnières muettes

$$D_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ \vdots \end{bmatrix} \quad D_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \\ \vdots \end{bmatrix} \quad D_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ \vdots \end{bmatrix},$$

que nous avons rencontrées pour la première fois dans la Section 1.4. Ces variables muettes ont été définies de telle sorte que leur somme est nulle

pour une année. Supposons maintenant que l'on régresse un vecteur à  $n$  composantes  $\mathbf{y}$  sur une constante et sur  $\mathbf{D} \equiv [\mathbf{D}_1 \ \mathbf{D}_2 \ \mathbf{D}_3]$ :

$$\mathbf{y} = \beta + \mathbf{D}\boldsymbol{\gamma} + \mathbf{u}. \quad (19.33)$$

Alors une série  $\mathbf{y}^*$  “ajustée par saison” peut être élaborée comme suit: as

$$\mathbf{y}^* \equiv \hat{\beta} + \hat{\mathbf{u}}, \quad (19.34)$$

où  $\hat{\beta}$  est l'estimation de  $\beta$ , et  $\hat{\mathbf{u}}$  est le vecteur de résidus provenant de l'estimation par OLS de (19.33). Ainsi toutes les variations de  $\mathbf{y}$  qui peuvent avoir comme explication des variables saisonnières muettes ont été éliminées pour construire  $\mathbf{y}^*$ .

Cette approche fut préconisée par Lovell (1963). Il montra, par une application du Théorème FWL, que les estimations OLS obtenues à partir des deux régressions suivantes étaient identiques:

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u} \quad \text{et} \quad (19.35)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\gamma} + \mathbf{u}. \quad (19.36)$$

Ici la première régression utilise des données “ajustées par saison” par la procédure utilisée en (19.33) et (19.34). La seconde se contente de régresser les données brutes  $\mathbf{y}$  sur des données brute  $\mathbf{X}$ , où  $\mathbf{X}$  doit contenir une constante ou un régresseur équivalent, et sur les variables saisonnières muettes  $\mathbf{D}$ . Ce résultat semble suggérer qu'il est peu important d'utiliser soit des données ajustées par saison soit des données brutes et des variables saisonnières muettes correspondant aux saisons. Une telle conclusion est exacte uniquement si les données ont été ajustées par saison à l'aide d'une régression.

Il existe de nombreux problèmes concernant l'ajustement saisonnier par régression. Premièrement, il est clair à partir des résultats standards sur les résidus des moindres carrés qu'avec des échantillons finis une régression comme (19.33) réduira la variation dans une trop grande mesure, en attribuant, à tort, la variation des variables saisonnières muettes (Thomas et Wallis, 1971). En second lieu, s'il existe une tendance croissante dans la série ajustée, une régression comme (19.33) attribuera à tort une partie de cette tendance aux variables saisonnières muettes. Par conséquent, l'estimation de l'effet du premier trimestre sera trop faible, et celle de l'effet du quatrième trimestre sera trop forte. Une solution évidente consiste à ajouter une tendance à la régression et à la traiter de la même manière qu'une constante. (Jorgenson, 1964). Cela implique, malgré tout, que  $\mathbf{X}$  doit inclure une tendance et une constante qsi l'on veut que (19.35) et (19.36) produisent en effet les mêmes estimations.

Le plus sérieux problème concernant l'approche de la régression et qu'elle ne permet pas de changement dans l'allure de la saisonnalité à travers le temps.

Comme la Figure 19.1 l'illustre, les llures saisonnières paraissent vraiment changer dans le temps. Une façon de modéliser ce phénomène consiste à ajouter des variables saisonnières muettes additionnelles qui ont été combinées à des puissances d'une tendance chronologiques annuelle linéaire croissante telle que

$$\mathbf{T} \equiv [1 \ 1 \ 1 \ 1 \ 2 \ 2 \ 2 \ 2 \cdots].$$

La raison qui veut que la tendance doive prendre cette forme relativement curieuse est que cela garantit toujours la nullité de la somme des variables de tendance muettes sur la totalité de chaque année, lorsque cette tendance est multipliée par les variables saisonnières muettes. Si l'on multipliait simplement les variables saisonnières muettes par une tendance ordinaire, cela ne serait plus le cas.

Le Théorème FWL s'applique aux régressions (19.35) et (19.36) quelle que soit la manière dont les variables muettes aient été définies. Ainsi on peut avoir

$$\mathbf{D} \equiv [\mathbf{D}_1 \ \mathbf{D}_2 \ \mathbf{D}_3 \ \mathbf{D}_1*\mathbf{T} \ \mathbf{D}_2*\mathbf{T} \ \mathbf{D}_3*\mathbf{T} \ \mathbf{D}_1*\mathbf{T}*\mathbf{T} \ \mathbf{D}_2*\mathbf{T}*\mathbf{T} \ \mathbf{D}_3*\mathbf{T}*\mathbf{T}].$$

Il y a trois ensembles de variables saisonnières muettes: celles qui sont les plus classiques et constantes dans le temps, celles qui sont combinées à une tendance linéaire et celles qui sont combinées à une tendance quadratique. Le fait de donner une tendance à des variables saisonnières muettes paraît quelquefois bien fonctionner avec des échantillon finis, dans le sens où elles semblent fournir une bonne approximation à un quelconque schéma courant de changement de saisonnalité. Mais cela n'a pas de sens asymptotiquement, parce que les variables saisonnières doivent en fin de compte devenir infinies si les coefficients associé aux variables de tendance muettes sont non nuls dans la régression.

En ce qui concerne les constructions de bâtiments sur la Figure 19.1, il est intéressant de voir que les variables de tendance muettes ne sont d'aucun usage. La régression de ces données sur une constante et trois variables saisonnières muettes produit quatre coefficients significatifs et un  $R^2$  d'environ 0.48. L'ajout de trois variables de tendance linéaire et trois variables de tendance quadratiques muettes à la régression n'améliore pas les valeurs ajustées de manière significative. Ainsi il apparaît, soit que la variation saisonnière de cette série n'a pas été modifiée dans le temps, malgré l'impression visuelles qu'elle donne, soit que cette modification s'est déroulée d'une manière telle qu'elle ne peut pas être approximée de façon satisfaisante par une régression sur des variables de tendance saisonnières muettes.

Un autre moyen de traiter les schéma saisonniers qui varient dans le temps consiste à utiliser les méthodes du domaine de fréquence; voir Engle (1974), Sims (1974), et Hylleberg (1977, 1986). La prmeière étape consiste à transformer les données  $y_t$  du domaine chronologique au domaine des fréquences,

habituellement à l'aide d'une transformation de Fourier.<sup>4</sup> Après transformation, chaque observation correspond à une certaine fréquence plutôt qu'à une certaine période de temps. Certaines observations sont effacées, en bandes autour des fréquences saisonnières et de leur harmonique. Le nombre d'observations effacées (c'est-à-dire les fréquences) est d'autant plus élevé que les bandes sont larges, et cela augmente la probabilité que toute variation saisonnière ait été éliminée des données. Enfin, les données sont transformées à nouveau pour aboutir dans le domaine chronologique, donnant une série ajustée par saison.

Sims (1974) montra que cette technique est équivalente à une forme d'ajustement saisonnier à l'aide d'une régression. Considérons la régression (19.33) et la série ajustée par saison définie par (19.34). Cette dernière serait équivalente à une série ajustée dans le domaine des fréquences que nous venons de décrire si la matrice  $\mathbf{D}$  était redéfinie de manière à être égale à un certain ensemble de variables qui sont des fonctions trigonométriques du temps. Les trois premières ou les onze premières de ces variables (dans le cas de données trimestrielles ou mensuelles respectivement) engendrent exactement le même sous-espace que trois ou onze variables saisonnières muettes. Ainsi si le schéma saisonnier était constant dans le temps, il serait nécessaire d'exclure seulement autant de fréquences spécifiques qu'il y a de périodes chronologiques dans l'année. L'exclusion de fréquences supplémentaires en bandes autour des fréquences saisonnières et de leur harmonique permet au schéma saisonnier de changer au cours du temps. Cela équivaut à inclure des fonctions trigonométriques du temps supplémentaires dans la régression. Le nombre de variables trigonométriques à inclure, qui est identique au nombre de fréquences exclues dans l'approche par le domaine des fréquences, augmentera de façon linéaire avec la taille de l'échantillon si la largeur des bandes demeure inchangée.

Le bureaux de statistiques officiels n'emploient presque jamais aucune sorte de procédure d'ajustement saisonnier basée sur la régression. Au delà des problèmes liés à de telles procédures et auxquels nous avons fait référence, elles souffrent d'une difficulté pratique importante. Au fur et à mesure que le temps passe et que la taille de l'échantillon s'accroît, l'estimation du vecteur  $\gamma$  dans (19.33) se modifie, et par conséquent chaque élément de  $\mathbf{y}^*$  sera modifié chaque fois qu'une nouvelle observation sera disponible. Cette caractéristique est à l'évidence la moins souhaitable pour les utilisateurs des statistiques officielles.

Les procédures d'ajustement saisonnier qui sont en réalité employées par les agences statistiques sont en général très compliquées. Elles tentent de traiter une multitude de problèmes pratiques, et parmi eux les tendances, les

<sup>4</sup> Pour une introduction aux méthodes du domaine de fréquence, consulter Harvey (1981). Pour une description de la transformation de Fourier, voir Press, Flannery, Teukolsky, et Vetterling (1986, Chapitre 12).



variations chronologiques des saisons, les variations du nombre de jours de commerce et les dates des vacances, le fait qu'une information plus pauvre caractérise le début de l'échantillon (parce que les observations qui précèdent l'échantillon sont inconnues), et les identités qui peuvent lier certaines séries entre elles. Ces procédures sont à l'origine conçues pour produire des données qui sont facilement lisibles par les économistes qui tentent de déterminer les performances de l'économie, plutôt que des données qui seront nécessairement plus utiles à des économètres. La plus connue de ces procédures officielles est la méthode du X-11 inventée par le Bureau de Recensement des Etats Unis (Shisken, Young, et Musgrave, 1967). Pour une discussion sur ce sujet et sur les procédures qui s'en inspirent, consulter Hylleberg (1986); la Figure 5.1 de cet ouvrage illustre le diagramme des opérations successives, qui révèle la complexité extrême de la procédure X-11.

Malgré la complexité de la procédure X-11 et de ses variantes, on peut souvent les approximer avec satisfaction par des procédures beaucoup plus simples basées sur ce que l'on appelle les **filtres linéaires**. Posons  $\mathbf{y}$  un vecteur à  $n$  composantes des observations (souvent en logarithmes plutôt qu'en niveaux) d'une série qui n'a pas été ajustée par saison. Un filtre linéaire est une matrice  $\Phi$  de dimension  $n \times n$  dont la somme des éléments d'une même ligne égale 1, qui prémultiplie  $\mathbf{y}$  pour produire une série ajustée par saison  $\mathbf{y}^*$ . Chaque ligne du filtre est un vecteur de **poïds filtrants**. Ainsi chaque élément de  $y_t^*$  de la série ajustée par saison est égal à une somme pondérée des valeurs passées, actuelle, et futures de  $y_t$ .

Considérons l'exemple simple de données trimestrielles. Supposons que l'on crée tout d'abord des moyennes mobiles à trois et onze termes

$$z_t \equiv \frac{1}{3}(y_{t-4} + y_t + y_{t+4}) \quad \text{et} \quad w_t \equiv \frac{1}{11} \sum_{j=5}^{-5} y_{t-j}.$$

La différence entre  $z_t$  et  $w_t$  est une estimation mobile de la quantité par laquelle la valeur de  $y_t$  du trimestre en cours tend à différer de sa valeur moyenne sur l'année. Ainsi une manière de définir une série ajustée par saison serait d'écrire

$$\begin{aligned} y_t^* &\equiv y_t - z_t + w_t \\ &= .0909y_{t-5} - .2424y_{t-4} + .0909y_{t-3} + .0909y_{t-2} \\ &\quad + .0909y_{t-1} + .7576y_t + .0909y_{t+1} + .0909y_{t+2} \\ &\quad + .0909y_{t+3} - .2424y_{t+4} + .0909y_{t+5}. \end{aligned} \tag{19.37}$$

Cet exemple correspond à un filtre linéaire dans lequel la ligne  $p$  de  $\Phi$  (pour  $5 < p < n - 5$ ) serait composée de  $p - 6$  zéros, suivis par onze coefficients qui apparaissent dans (19.37), eux-mêmes suivis par  $n - p - 5$  zéros.

Cet exemple tel qu'il fut construit fut délibérément trop simple, mais l'approche de base qu'il illustre se retrouve, sous des formes modifiées variées,

dans la plupart des procédures d'ajustement saisonnier officielles. Ces dernières n'emploient généralement pas des filtres linéaires, mais plutôt des moyennes mobiles sous une forme comparable à cet exemple. Ces moyennes mobiles tendent à être plus longues que celles utilisées dans l'exemple;  $z_t$  est généralement composée d'au moins 5 termes et  $w_t$  d'au moins 25 termes dans le cas de données trimestrielles. Elles tendent également à donner progressivement moins de poids aux observations éloignées de  $t$ . Le poids donné à  $y_t$  par ces procédures est généralement compris entre 0.75 et 0.9, mais il est toujours inférieur à 1. Pour plus de détails sur les relations entre les procédures officielles et celles basées sur les filtres linéaires, voir Wallis (1974), BurrIDGE et Wallis (1984), et Ghysels et Perron (1993).

Nous avons affirmé que les procédures d'ajustement saisonnier officielles ont les mêmes propriétés la plupart du temps que les filtres linéaires appliqués soit aux niveaux soit aux logarithmes des données brutes. Cette assertion peut être vérifiée empiriquement. Si elle est exacte, la régression d'une série ajustée par saison  $y_t^*$  sur suffisamment de retards et d'avances de la série brute correspondante  $y_t$  devrait fournir des valeurs ajustées d'une qualité extrême. Le coefficient de  $y_t$  devrait être élevé et positif, mais inférieur à 1, et les coefficients des  $y_{t+j}$  devraient être négatifs lorsque  $j$  est un entier multiple de 4 ou de 12, pour des données trimestrielles et mensuelles respectivement.

Pour illustrer ces propos, nous avons régressé les logarithmes de la série des constructions de bâtiments pour le Canada ajustée par saison qui correspond à la série brute de la Figure 19.1 sur une constante et sur la valeur courante et 12 retards et avances de la série brute, pour la période allant de 1957:1 à 1986:4. Le  $R^2$  est de .992 et le coefficient de la valeur courante est de 0.80. Nous avons également régressé les logarithmes des dépenses de consommations réelles des particuliers, ajustées par saison, sur une constante, la valeur courante et 13 retards et avances de la série brute correspondant, pour la période allant de 1953:1 à 1984:1.<sup>5</sup> Cette fois-ci, le  $R^2$  atteint la valeur extraordinaire de .999996, et le coefficient associé à la valeur courante est 0.82. Dans les deux cas, tous les coefficients associés à  $y_{t+j}$  pour  $j$  un multiple de 4 étaient négatifs, comme prévu. Il apparaît donc qu'un filtre linéaire fournit une approximation de grand qualité de la procédure d'ajustement saisonnier employée en réalité dans le cas de données de dépenses et une approximation satisfaisante dans le cas des données de construction de bâtiments.

Si l'on réalise un ajustement saisonnier à l'aide d'un filtre linéaire, il n'est pas difficile d'analyser les effets de l'utilisation de données ajustées par saison. Supposons que le *même* filtre soit appliqué à toutes les séries dans la régression

<sup>5</sup> Toutes les données furent collectées à partir de la banque de données CANSIM des Statistiques Canadiennes. Les séries de construction de bâtiments ajustées et brutes portent les numéros D2717 et D4945. Les séries des dépenses ajustées et brutes portent les numéros D20131 et D10131.

de  $\mathbf{y}^*$  sur  $\mathbf{X}^*$ . Alors les estimations par moindres carrés seront données par

$$\begin{aligned}\tilde{\beta} &= (\mathbf{X}^{*\top} \mathbf{X}^*)^{-1} \mathbf{X}^{*\top} \mathbf{y}^* \\ &= (\mathbf{X}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{y}.\end{aligned}$$

Nous voyons que  $\tilde{\beta}$  est simplement un vecteur d'estimations GLS, où la matrice de dimension  $n \times n$   $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$  joue le rôle de l'inverse de la matrice de covariance des aléas. Nous concluons donc que la régression OLS suivant l'ajustement saisonnier pratiqué à l'aide d'un filtre linéaire est équivalent à une régression GLS, à condition que le même filtre linéaire soit employé pour toutes les séries. Malheureusement, les procédures d'ajustement saisonnier ne pratiquent pas ainsi pour toutes les séries (ni quelquefois pour une même série en différents points du temps). Par conséquent, ce résultat est rarement applicable. (Wallis, 1974).

Quoi qu'il en soit, il y a un intérêt à discuter des propriétés de  $\tilde{\beta}$ . Celles-ci dépendront à l'évidence de la manière dont on a généré  $y_t$ . L'une des possibilités est que

$$\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (19.38)$$

qui implique que n'importe quelle forme de saisonnalité dans  $\mathbf{y}$  soit rendue dans sa totalité par la saisonnalité dans les variables indépendantes. Alors

$$\text{plim}_{n \rightarrow \infty} \tilde{\beta} = \beta_0 + \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{X} \right)^{-1} \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \mathbf{X}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \mathbf{u} \right) = \beta_0. \quad (19.39)$$

Ainsi, bien qu'il n'y ait aucune raison d'utiliser des données ajustées par saison dans ce cas, leur présence conserve quand même la convergence des estimations par moindres carrés. Cependant, le Théorème de Gauss-Markov implique que ces estimations seront moins efficaces que les estimations OLS qui utilisent les données brutes. C'est le cas, puisque la procédure d'ajustement saisonnier réduit la variation des variables indépendantes et elle réduit également la précision de l'estimation de  $\beta$ . Ce plus, la seconde égalité de (19.39) réclame que tous les éléments de  $\mathbf{X}$  soient indépendants de tous les éléments de  $\mathbf{u}$ , et elle élimine implicitement la possibilité d'inclure des variables dépendantes retardées dans la matrice  $\mathbf{X}$ .

Une seconde possibilité, qui rend l'utilisation de données ajustées par saison plus attrayante est que le DGP soit

$$\mathbf{y} - \mathbf{y}^s = (\mathbf{X} - \mathbf{X}^s)\beta_0 + \mathbf{u}, \quad \mathbf{u} \sim \text{IID}(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (19.40)$$

Ici  $\mathbf{y}^s$  et  $\mathbf{X}^s$  désignent les parties de  $\mathbf{y}$  et  $\mathbf{X}$  attribuées aux saisons. Supposons que les poids filtrants aient été choisis de telle manière que toute saisonnalité

soit éliminée. Cela implique que  $\Phi \mathbf{y}^s = \mathbf{0}$  et  $\Phi \mathbf{X}^s = \mathbf{0}$ , ce qui implique en retour que

$$\begin{aligned}\Phi \mathbf{y} &= \Phi((\mathbf{X} - \mathbf{X}^s)\beta_0 + \mathbf{y}^s + \mathbf{u}) \\ &= \Phi(\mathbf{X}\beta_0 + \mathbf{u}).\end{aligned}$$

Si l'on substitue  $\Phi(\mathbf{X}\beta_0 + \mathbf{u})$  à  $\Phi \mathbf{y}$  dans la première ligne de (19.39), sans changer la suite de (19.39), on conclue que  $\tilde{\beta}$  est convergent vers  $\beta_0$ .

Dans cette seconde situation, l'alternative consistant simplement à régresser les données brutes  $\mathbf{y}$  sur  $\mathbf{X}$  n'est pas du tout attrayante. L'estimation OLS de  $\beta$  est

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= \beta_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (-\mathbf{X}^s \beta_0 + \mathbf{y}^s + \mathbf{u}),\end{aligned}$$

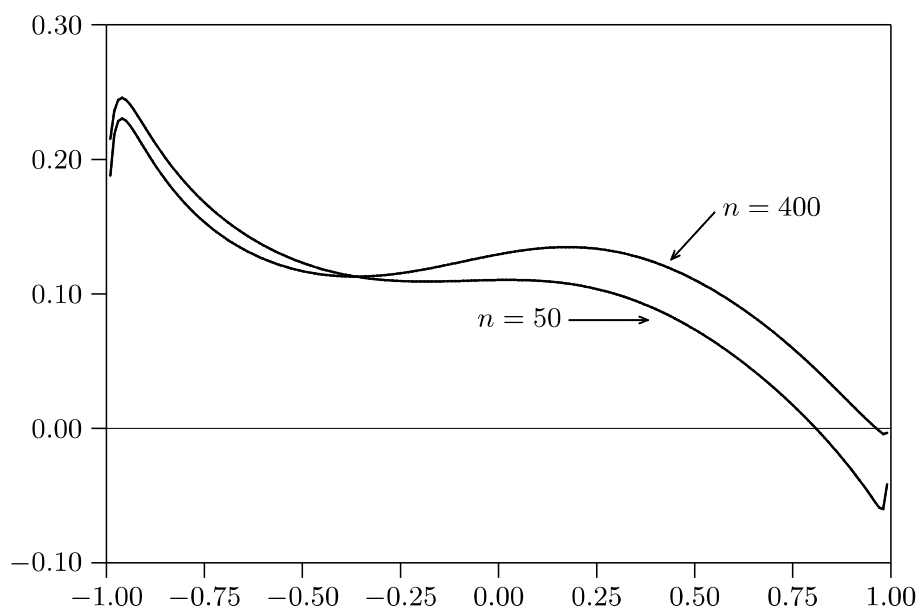
et elle ne sera bien évidemment pas convergente vers  $\beta_0$  à moins que  $\mathbf{X}$  ne soit asymptotiquement orthogonale à la fois à  $\mathbf{X}^s$  et  $\mathbf{y}^s$ . Mais une telle condition ne peut être valide que si aucune variable de  $\mathbf{X}$  ne manifeste une quelconque variation saisonnière. Par conséquent, si l'on désire utiliser des données ajustées par saison, il faut incorporer une saisonnalité de façon explicite dans le modèle. Nous traiterons ce thème dans la section qui suit.

Souvenons-nous que ces résultats ne sont valides que si le même filtre linéaire est utilisé pour l'ajustement saisonnier de toutes les séries. Si l'on multiplie les filtres pour les différentes séries, ce qui sera presque toujours le cas avec des données ajustées par les procédures officielles, on ne peut plus affirmer que les régressions qui emploient des données ajustées par saison produiront des estimations convergentes, que les données aient été générées par un modèle comme (19.38) ou par un modèle comme (19.40). On peut juste espérer qu'une telle défaillance dans la convergence soit faible. Consulter Wallis (1974).

Une limitation beaucoup plus sérieuse concernant la convergence dans les résultats précédents est qu'il supposent l'absence totale de variable dépendante retardée parmi les régresseurs. Lorsqu'il existe de telles variables, et cela sera le cas pour tout modèle dynamique et pour tout modèle transformé de façon à permettre la corrélation en série des aléas, il n'y a aucune raison de croire que la régression par moindres carrés utilisant des données ajustées avec un filtre linéaire produira des estimations convergentes. En réalité, des travaux récents ont montré que, dans les modèles comportant un seul retard de la variable dépendante, l'estimation du coefficient de la variable retardée tend généralement à être sévèrement biaisé lorsque l'on utilise des données ajustées par saison. Consulter Jaeger et Kunst (1990), Ghysels (1990), et Ghysels et Perron (1993).

Afin d'illustrer ce résultat important, nous avons généré des données artificielles à partir d'un cas particulier du modèle

$$y_t = \alpha + \beta y_{t-1} + \mathbf{D}_t \gamma + u_t, \quad u_t \sim N(0, \sigma^2), \quad (19.41)$$



**Figure 19.2** Biais dû à l'ajustement saisonnier

où  $D_t$  est la  $t^{\text{ième}}$  ligne d'une matrice de dimension  $n \times 3$  de variables saisonnières muettes. La série  $y_t$  a ensuite été soumise à un filtre linéaire que l'on pourrait utiliser pour l'ajustement saisonnier,<sup>6</sup> et la série "ajustée" a ensuite été régressée sur une constante et sur sa propre valeur retardée pour fournir une estimation  $\tilde{\beta}$ . Nous avons exécuté cette procédure pour 199 valeurs de  $\beta$  allant de  $-0.99$  à  $0.99$ , pour des tailles d'échantillons diverses, et les expériences furent répétées un grand nombre de fois afin de réduire l'erreur expérimentale (voir le Chapitre 21).

La Figure 19.2 illustre le biais estimé de  $\tilde{\beta}$  en fonction de  $\beta$ . Seuls les résultats pour  $n = 50$  (basé sur 4000 exécutions) et pour  $n = 400$  (basés sur 2000 exécutions) sont reportés. Remarquons que  $n$  est le nombre des observations pour les séries ajustées par saison, qui est inférieur de 54 au nombre des observations initiales. On voit clairement à partir de la figure que, pour la plupart des valeurs de  $\beta$ ,  $\tilde{\beta}$  est sévèrement biaisé vers le haut. Ce biais ne se dissipe pas lorsque la taille de l'échantillon s'accroît; en réalité, pour de nombreuses valeurs de  $\beta$ , il est plus fort avec  $n = 400$  qu'avec  $n = 50$ . La conclusion semble inéluctable que  $\tilde{\beta}$  est un estimateur non convergent et que l'amplitude de cette non convergence est en général assez forte.

Un autre résultat intéressant est ressorti de cette batterie d'expériences. L'estimation de  $\sigma$  qui utilise les données ajustées par saison est biaisée vers

<sup>6</sup> La valeur courante de la série brute est associée au poids 0.84. Les 12 valeurs de retard et d'avance sont associées aux poids 0.08, 0.07, 0.06,  $-0.16$ , 0.05, 0.05, 0.04,  $-0.12$ , 0.03, 0.03, 0.02, et  $-0.08$ . Les valeurs particulières de ces poids n'ont pas affecté les résultats qualitatifs.

le bas dans une large mesure, avoisinant en moyenne entre 87% et 92% de sa véritable valeur. Par contre, lorsque le modèle exact (19.41) est estimé à l'aide des données brutes, l'estimation de  $\sigma$  est pratiquement sans biais, comme prévu. Ces résultats convergent vers les résultats obtenus par Plosser (1979a), qui trouva que les modèles estimés avec des données ajustées par saison possèdent toujours des variances de résidus plus faibles que celles correspondant aux modèles estimés avec les données brutes. Quoi qu'il en soit, Plosser trouva que les prévisions fondées sur ces derniers seront plus fines que celles fondées sur les premiers. Ces conclusions suggèrent que l'on ne devrait jamais choisir un modèle basé sur les données ajustées par saison plutôt qu'un modèle basé sur les données brutes simplement parce que les premiers semblent s'ajuster un peu mieux.

L'usage des données ajustées par saison dans les travaux économétriques appliqués est très répandu, et il est en vérité quelquefois difficile de l'éviter. Cependant les résultats exposés dans cette section suggèrent que cette attitude peut souvent être imprudente. Même pour des modèles statiques, il est probable que des problèmes surgissent si les procédures officielles d'ajustement saisonnier utilisent en réalité des filtres différents. Pour les modèles dynamiques la non convergence potentielle provenant de l'utilisation de données ajustées par saison paraît très marquée. Dans la prochaine section, nous discuterons par conséquent des approches variées de la spécification et de l'estimation des modèles qui emploient des données qui ne sont pas ajustées par saison.

## 19.7 MODÉLISER LA SAISONNALITÉ

Les résultats de la section qui précède suggèrent que, lorsque l'on dispose des données brutes, il est probablement plus judicieux de les utiliser plutôt que de s'appuyer sur des données officielles ajustées par saison. Malgré tout, cela réclame une bonne quantité de travail supplémentaire. L'estimation simple d'un modèle qui n'est pas conçu pour des données saisonnières est rarement appropriée. Une telle approche a toutes les chances de produire des estimations des paramètres sévèrement biaisées si la variation saisonnière d'une ou de plusieurs variables indépendantes s'avère être corrélée (même si elle ne la provoque pas) avec la variation saisonnière de la variable dépendante. Il existe de nombreux moyens de gérer la variation saisonnière dans les modèles de régression. C'est dans cette section que nous discutons de certaines d'entre elles.

La stratégie la plus simple pour la spécification de modèles qui utilisent des données brutes consiste à inclure des variables saisonnières muettes dans le modèle de régression linéaire, comme dans (19.36). Si la structure saisonnière a été constante au cours du temps, de sorte que les trois variables saisonnières muettes (dans le cas de données trimestrielles) ou les onze variables saisonnières

muettes (dans le cas de données mensuelles) rendent compte de façon satisfaisante des effets de la saisonnalité, cette approche semble être adéquate. Cependant, elle ne sera pas appropriée lorsque la structure de la saisonnalité des variables dépendantes ou indépendantes est changeante au cours de la période d'échantillonnage. Une possibilité dans ce cas consiste à inclure un ou plusieurs ensembles de variables saisonnières muettes combinées à des tendances annuelles croissantes, en même temps que des variables saisonnières muettes ordinaires. La pertinence des ensembles additionnels de variables muettes peut facilement être testée aux moyens des tests en  $F$  à la manière habituelle. Une critique à cette approche, ainsi que nous l'avons noté précédemment, est qu'elle n'a pas de sens asymptotiquement. De plus, un modèle qui possède des variables saisonnières à tendance à toutes les chances d'être inadapté à la prévision, puisque même si les variables saisonnières muettes rendent compte de façon satisfaisante des changements de la structure de la saisonnalité dans l'échantillon, il n'y a pas de raison de croire qu'elles le feront en dehors de l'échantillon. Davidson et MacKinnon (1983c) offrent un exemple quelque peu extrême de cette approche. Dans cet article, pas moins de 15 variables saisonnières muettes, avec des tendances allant jusqu'au quatrième ordre, furent incluses dans des modèles utilisant des données trimestrielles, parce que cela semblait être nécessaire pour rendre compte de toute la saisonnalité dans les données.

Une seconde stratégie consiste à modéliser les aléas d'un modèle de régression pour qu'ils obéissent à une espèce quelconque de **processus ARMA saisonnier**, c'est-à-dire un processus ARMA avec des coefficients non nuls uniquement sur les retards des saisons. Un tel processus, qui peut être adéquat pour les données trimestrielles, est le processus Ar(1) simple que nous avons rencontré pour la première fois dans la Section 10.5:

$$u_t = \rho_4 u_{t-4} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (19.42)$$

où  $\rho_4$  est le paramètre à estimer, et  $\omega^2$  est la variance de  $\varepsilon_t$ . Un autre processus Ar purement saisonnier consacré aux données trimestrielles est

$$u_t = \rho_4 u_{t-4} + \rho_8 u_{t-8} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2), \quad (19.43)$$

qui est l'analogue d'un processus AR(2) consacré à des données non saisonnières.

Dans de nombreux cas, les aléas peuvent manifester à la fois de la corrélation saisonnière et de la corrélation non saisonnière. Cela suggère que l'on peut combiner un processus saisonnier avec un processus qui ne l'est pas. Supposons, par exemple, que l'on veuille combiner un processus AR(1) avec un processus AR(4) simple. Une approche ferait combiner ces deux processus de façon additive, produisant

$$u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (19.44)$$

Une seconde approche ferait combiner ces deux processus de façon multiplicative, comme dans

$$(1 - \rho_1 L)(1 - \rho_4 L^4)u_t = \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2),$$

que l'on pourrait écrire différemment, en oubliant la notation avec l'opérateur retard, comme dans

$$u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} - \rho_1 \rho_4 u_{t-5} + \varepsilon_t, \quad \varepsilon_t \sim \text{IID}(0, \omega^2). \quad (19.45)$$

Aussi bien (19.44) que (19.45) paraissent probables, et il n'existe aucune raison majeure a priori de préférer l'un à l'autre.

A l'évidence, un grand nombre de processus AR et ARMA différentes pourraient être employés pour modéliser la variation saisonnière de l'aléa dans un modèle de régression. Il existe une littérature très développée sur les processus ARMA saisonniers; consulter, parmi d'autres auteurs, Box et Jenkins (1976), Harvey (1981), et Ghysels (1991). Cependant, l'intérêt que représentent de tels processus pour modéliser la saisonnalité n'est pas de tout immédiat. D'un côté, ils offrent généralement une façon assez économe de le faire; par exemple (19.42) n'emploie qu'un seul paramètre additionnel, et (19.13) n'en a que deux. De plus, il est certainement exact que si un modèle de régression ne rend pas compte de façon adéquate de la saisonnalité, la corrélation sérielle d'ordre quatre se manifesterait nécessairement. Alors le test de cette corrélation fournit souvent un test diagnostique utile. Mais, de même que la corrélation en série à l'ordre un ne signifie pas que les aléas obéissent en vérité à un processus AR(1), la corrélation en série à l'ordre quatre ne signifie pas non plus qu'ils obéissent à un processus AR(4).

L'énorme difficulté relative aux processus ARMA saisonniers est qu'ils ne peuvent pas saisir l'un des caractéristiques importantes de la saisonnalité, en l'occurrence le fait que des saisons différentes de l'année possèdent des particularités différentes: l'été n'est pas simplement l'hiver avec un nouveau nom. Mais en ce qui concerne un processus ARMA, l'été *est* juste l'hiver avec un nom différent. Si les aléas obéissent à un schéma saisonnier particulier au début de l'échantillon, alors il est assez probable qu'ils obéissent au même schéma l'année suivante. Mais pour un processus ARMA stationnaire, l'influence des conditions initiales tend vers zéro lorsque le temps passe. Ainsi il n'y a aucune raison de croire que le schéma saisonnier 10 ou 20 ans après le début de l'échantillon possèdera une quelconque ressemblance avec le schéma d'origine. En fait, pour  $T$  suffisamment élevé, les espérances de  $u_T$ ,  $u_{T+1}$ ,  $u_{T+2}$ , et  $u_{T+3}$  conditionnellement à  $u_1$ ,  $u_2$ ,  $u_3$  et  $u_4$  sont toutes (presque) nulles. Alors l'utilisation d'un processus ARMA pour modéliser la saisonnalité implique l'hypothèse que tout schéma de saisonnalité particulier est transitoire; dans le long terme, tout schéma est envisageable. Cela nous entraîne à croire que l'on utilisera sûrement pas le schéma saisonnier ARMA pour modéliser le schéma saisonnier d'un objet tel que le prix des framboises, puisque le



modèle serait incapable d'expliquer que le prix a toutes les chances d'être inhabituellement élevé au milieu de l'hiver ou lors de la récolte. Un moyen évident de contourner ce problème serait d'inclure des variables saisonnières muettes dans le modèle. Les variables saisonnières muettes permettraient aux différentes saisons d'être naturellement différentes, alors que le processus ARMA saisonnier permettrait au schéma saisonnier d'évoluer dans le temps.

Une troisième stratégie consiste à permettre à certains coefficients de la fonction de régression de varier dans chaque saison. Ainsi, si le modèle originel possède  $k$  coefficients, on estimerait un modèle avec  $4k$  ou  $12k$  coefficients. Cela serait pertinent si les variations du schéma de saisonnalité dans le temps étaient associées à des modifications des valeurs de certaines variables indépendantes dans le temps. Une objection immédiate à cette approche est que le nombre de coefficients serait souvent très élevé comparativement à la taille de l'échantillon, et ils tendront tous à être estimés avec trop peu de précision. Gersovitz et MacKinnon (1978) ont à cette occasion suggéré l'utilisation des informations a priori de régularité, comparables à celles dont nous avons discuté lors de la Section 19.3 pour l'estimation des retards échelonnés, afin d'éviter des variations trop fortes des coefficients d'une saison à l'autre. Cela paraît être une contrainte raisonnable à imposer dans le cas de données mensuelles, mais cela paraîtrait difficile à justifier dans le cas de données trimestrielles;

Une quatrième stratégie consiste à incorporer des dynamiques saisonnières directement dans la spécification de la fonction de régression, à l'aide d'une forme quelconque de **modèle ADL saisonnier**. Un modèle particulièrement simple de ce genre est

$$(1 - L^4)y_t = \beta_0 + \beta_1(1 - L^4)x_t + \beta_2(y_{t-4} - \lambda x_{t-4}) + u_t.$$

Cela ressemble à un modèle ADL(1,1) écrit sous sa forme à correction d'erreur — à comparer à (19.30) — mais avec des retards à la quatrième période au lieu des retards à une période. Il est presque certainement trop simple, bien sûr, et l'addition de variables saisonnières muettes ou de retards de  $y_t$  et  $x_t$ . Un article très connu qui estime les modèles ALD saisonniers fut écrit par Davidson, Hendry, Srba, et Yeo (1978).

A l'exception discutable des modèles ADL saisonniers, les stratégies aperçues jusqu'à présent sont essentiellement mécaniques. On commence avec un modèle non saisonnier et on le transforme afin de lui faire manipuler la saisonnalité. Ce n'est sûrement pas le meilleur moyen de procéder. Dans un monde idéal, on aimerait incorporer la saisonnalité dès le départ dans le modèle. Cela a pourtant toutes les chances de rendre l'élaboration du modèle beaucoup plus difficile, et cela explique sans doute pourquoi peu d'auteurs s'y sont attaqués, à l'exception de Plosser (1979b), Miron (1986), et Osborn (1988, 1991). A moins que la théorie économique ne prenne explicitement en compte la saisonnalité, il sera très difficile aux économètres d'intégrer cette saisonnalité dans les modèles qu'ils estiment.

## 19.8 CONCLUSION

Dans ce chapitre, nous avons vu un certain nombre de problèmes qui apparaissent fréquemment lorsque l'on tente d'estimer des modèles de régression à l'aide de données temporelles. Dans la majeure partie du chapitre, nous avons supposé que toutes les séries sont stationnaires, ou  $I(0)$ , de sorte que l'on peut employer des méthodes d'estimation classiques et la théorie asymptotique standard. Pour de nombreuses séries cependant, cette hypothèse peut être enfreinte à moins de prendre les différences premières avant l'estimation. Mais comment sait-on qu'une opération des différences premières est nécessaire? Dans le chapitre qui suit, nous discutons de la manière de répondre à cette question, et il nous permet d'aborder des thèmes importants qui lui sont rattachés.

## TERMES ET CONCEPTS

retards d'Almon	modèle d'ajustement partiel
modèles $ADL(p, q)$ et modèles $ADL(1, 1)$	modèle $PDL(q, d)$
modèle autorégressif à retard échelonné (ADL)	retard échelonné polynomial (PDL)
modèle dynamique	marche aléatoire, avec ou sans dérive
forme à correction d'erreur (d'un modèle ADL)	procédure d'ajustement saisonnier basée sur la régression
terme de correction d'erreur	saisonnalité
poids filtrant	données ajustées par saison
opérateur de la différence première	modèle ADL saisonnier
causalité au sens de Granger dans les VAR	processus AR saisonnier
variables intégrées	variation saisonnière
variables $I(0)$ et $I(1)$	informations a priori de régularité
filtre linéaire	régression erronée
estimation mixte	contraintes stochastiques
	variable à tendance stationnaire
	modèle à vecteur autorégressif
	processus $VAR(p)$

# Chapitre 20

## Racines Unitaires et Cointégration

### 20.1 INTRODUCTION

Comme nous l'avons vu dans le chapitre précédent, on ne peut pas s'attendre à ce que les résultats asymptotiques s'appliquent si une quelconque variable dans un modèle de régression est générée par un processus non stationnaire. Par exemple, dans le cas du modèle de régression linéaire  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , les résultats habituels dépendent de l'hypothèse selon laquelle la matrice  $n^{-1}\mathbf{X}^\top \mathbf{X}$  tend vers une matrice finie, définie positive lorsque la taille de l'échantillon  $n$  tend vers l'infini. Lorsque cette hypothèse n'est pas vérifiée, des phénomènes extrêmement étranges peuvent survenir, comme nous l'avons vu lors de notre discussion dans la Section 19.2 sur les régressions "erronées" entre des variables sans aucune relation. Cela constitue un problème pratique sérieux, dans la mesure où un grand nombre de séries temporelles manifestent une tendance croissante à travers le temps, et semblent par conséquent enfreindre cette hypothèse.

Les deux moyens qui permettent de conserver l'hypothèse valide lorsque l'on emploie de telles séries consistent à éliminer la tendance ou à calculer les différences premières avant de les manipuler. Mais l'élimination de la tendance et le calcul des différences premières sont en réalité des opérations radicalement opposées: si la première est appropriée, la seconde ne l'est pas, et vice versa. Éliminer la tendance d'une série temporelle  $y_t$  sera pertinent si elle est stationnaire autour d'une tendance, ce qui implique que l'on peut écrire le DGP pour  $y_t$  sous la forme

$$y_t = \gamma_0 + \gamma_1 t + u_t, \quad (20.01)$$

où  $t$  est une tendance temporelle et où  $u_t$  obéit à un processus ARMA stationnaire. Alternativement, le calcul des différences sera pertinent lorsque le DGP pour  $y_t$  peut s'écrire sous la forme

$$y_t = \gamma_1 + y_{t-1} + u_t, \quad (20.02)$$

où  $u_t$  suit également un processus ARMA stationnaire. Si les  $u_t$  étaient non autocorrélés, (20.02) serait une marche aléatoire avec dérive, le paramètre de

dérive étant  $\gamma_1$ . Quoi qu'il en soit, les aléas seront autocorrélés, en général. Comme nous le verrons prochainement, le fait que le paramètre  $\gamma_1$  apparaisse à la fois dans (20.01) et (20.02) ne relève absolument pas du hasard.

Le choix entre l'élimination de la tendance et le calcul des différences se ramène à un choix entre (20.01) et (20.02). Les principales techniques de choix entre les deux sont des tests variés de ce que l'on appelle les **racines unitaires**. La terminologie provient de la littérature consacrée aux processus de séries temporelles. Souvenons-nous à partir de la Section 10.7 que pour un processus AR  $A(L)u_t = \varepsilon_t$ , où  $A(L)$  désigne un polynôme en l'opérateur retard, la stationnarité du processus dépend des racines de l'équation polynomiale  $A(L) = 0$ . Si toutes les racines sont à l'extérieur du cercle unitaire, le processus est stationnaire. Si une quelconque racine est égale ou inférieure à 1 en valeur absolue, le processus est non stationnaire. Une racine égale à 1 en valeur absolue est appelée **racine unitaire**. Lorsqu'un processus possède une racine unitaire, comme c'est le cas pour (20.02), on parle de processus **intégré d'ordre un** ou **I(1)**. Pour qu'une série  $I(1)$  soit stationnaire, il faut calculer ses différences premières.

Le moyen évident de choisir entre (20.01) et (20.02) consiste à les emboîter pour obtenir un modèle beaucoup plus général. Il existe un grand nombre de façons de procéder. Le modèle qui engloberait à la fois (20.01) et (20.02) de la façon la plus plausible serait

$$\begin{aligned} y_t &= \gamma_0 + \gamma_1 t + v_t; \quad v_t = \alpha v_{t-1} + u_t \\ &= \gamma_0 + \gamma_1 t + \alpha(y_{t-1} - \gamma_0 - \gamma_1(t-1)) + u_t, \end{aligned} \quad (20.03)$$

où  $u_t$  obéirait à un processus stationnaire. Ce modèle fut préconisé par Bhargava (1986). Lorsque  $|\alpha| < 1$ , (20.03) est équivalent au modèle stationnaire autour d'une tendance (20.01); lorsque  $\alpha = 1$ , il devient (20.02).

Parce que (20.03) est non linéaire en ses paramètres, il est commode de le reparamétriser comme suit

$$y_t = \beta_0 + \beta_1 t + \alpha y_{t-1} + u_t, \quad (20.04)$$

où

$$\beta_0 \equiv \gamma_0(1 - \alpha) + \gamma_1 \alpha \quad \text{et} \quad \beta_1 \equiv \gamma_1(1 - \alpha).$$

Il est aisé de vérifier que les estimations par moindres carrés de  $\alpha$  dans (20.03) et (20.04) seront identiques, tout comme les écarts types estimés de ces estimations si, dans le cas de (20.03), ces derniers sont basés sur la régression de Gauss-Newton. Le seul inconvénient de la reparamétrisation de (20.04) est qu'elle passe entièrement sous silence le fait que  $\beta_1 = 0$  lorsque  $\alpha = 1$ .

Si l'on retranche  $y_{t-1}$  aux deux membres, l'équation (20.04) devient

$$\Delta y_t = \beta_0 + \beta_1 t + (\alpha - 1)y_{t-1} + u_t, \quad (20.05)$$

où  $\Delta$  est l'opérateur des différences premières. Si  $\alpha < 1$ , (20.05) est équivalent au modèle (20.01), alors que si  $\alpha = 1$ , il est équivalent à (20.02). Ainsi il est habituel de tester l'hypothèse nulle  $\alpha = 1$  contre l'hypothèse alternative unilatérale  $\alpha < 1$ . Puisqu'il s'agit de tester l'hypothèse nulle de présence d'une racine unitaire dans le processus qui génère  $y_t$ , on appelle communément ces tests des **tests de racine unitaire**.

A première vue, il semblerait qu'un test de racine unitaire puisse être exécuté en observant simplement le  $t$  de Student ordinaire pour  $\alpha - 1 = 0$  dans (20.05), mais il n'en est rien. Lorsque  $\alpha = 1$ , le processus qui génère  $y_t$  est intégré d'ordre un. Cela signifie que  $y_{t-1}$  ne satisfera pas les hypothèses standards nécessaires à l'analyse asymptotique. En conséquence, comme nous allons le voir bientôt, le  $t$  de Student n'est pas asymptotiquement distribué suivant une  $N(0, 1)$ . On utilise en fait cette statistique comme un  $t$  de Student habituel, mais on ne l'associe pas aux valeurs critiques usuelles des distributions de Student ou normale.

La première moitié de ce chapitre est consacrée aux tests de racines unitaires. Dans la prochaine section, nous décrivons un certain nombre de tests de racines unitaires largement diffusés, tous étant basés sur des régressions comparables à (20.05), et reposant sur l'hypothèse peu réaliste que les aléas  $u_t$  ne sont pas autocorrélés. Dans la Section 20.3, nous discutons ensuite de quelques aspects de la théorie asymptotique qui s'est développée pour ces tests. Dans la Section 20.4, nous abandonnons l'hypothèse de non autocorrélation des aléas et discutons d'autres problèmes qui compliquent l'usage des tests de racines unitaires.

La seconde moitié du chapitre traite du concept fondamental de **co-intégration** entre deux ou plusieurs séries, chacune étant  $I(1)$ . Ce concept est introduit dans la Section 20.5. Les tests de cointégration, qui sont étroitement reliés aux tests de racines unitaires, sont abordés dans la Section 20.6. Le fait que la variable dépendante dans un modèle de régression soit cointégrée avec un ou plusieurs régresseurs entraîne un certain nombre de conséquences importantes sur le type de modèle qu'il faudrait élaborer. Dans la Section 20.7, nous discutons des méthodes équation par équation pour l'estimation à l'aide de séries  $I(1)$ , et dans la Section 20.8, nous discutons des méthodes basées sur des autorégressions vectorielles.

## 20.2 TESTS DE RACINES UNITAIRES

Les tests de racines unitaires les plus simples et les plus largement utilisés furent développés par Fuller (1976) et Dickey et Fuller (1979). On se réfère habituellement à ces tests en tant que **tests de Dickey-Fuller**, ou **tests DF**. On trouvera chez Dickey, Bell, et Miller (1986) un exposé particulièrement brillant de ces tests. Les tests de Dickey-Fuller se basent sur des régressions telles que (20.05). Trois régressions comparables sont communément employées, (20.05)

étant la plus compliquée. Les deux autres sont

$$\Delta y_t = (\alpha - 1)y_{t-1} + u_t \quad \text{et} \quad (20.06)$$

$$\Delta y_t = \beta_0 + (\alpha - 1)y_{t-1} + u_t. \quad (20.07)$$

On peut dériver ces deux régressions exactement de la même manière que (20.05). La première, (20.06), est extrêmement contraignante, tellement contraignante qu'il est difficile d'imaginer que l'on puisse l'employer avec des séries temporelles économiques. Son seul avantage est qu'elle est plus facile à analyser que les deux autres régressions. La seconde, (20.07), est également assez contraignante, mais elle serait intéressante si  $y_t$  ne possédait aucune tendance. Remarquons que, dans le cas de (20.07),  $\beta_0 = 0$  dès lors que  $\alpha = 1$ , parce que  $\beta_0$  est en fait  $\gamma_0(1 - \alpha)$ .

Il existe deux types distincts de tests DF basés sur chacune des trois régressions (20.05), (20.06), et (20.07). Un type de test est calculé exactement comme un  $t$  de Student ordinaire pour  $\alpha - 1 = 0$  dans n'importe quelle régression. Puisque ces statistiques ne suivent pas une distribution de Student, même asymptotiquement, on les nomme habituellement **statistiques  $\tau$**  plutôt que  $t$ . Nous nommerons les statistiques  $\tau$  basées sur (20.06), (20.07), et (20.05):  $\tau_{nc}$ ,  $\tau_c$ , et  $\tau_{ct}$ , respectivement.<sup>1</sup> Le second type de tests se base directement sur l'estimation du coefficient  $\hat{\alpha} - 1$ . La statistique de test est

$$z = n(\hat{\alpha} - 1). \quad (20.08)$$

Par analogie avec les trois statistiques  $\tau$ , nous noterons  $z_{nc}$ ,  $z_c$ , et  $z_{ct}$  les trois versions principales de la **statistique  $z$** .

La statistique  $z$  (20.08) peut paraître étrange pour deux raisons: elle ne dépend pas d'une estimation de  $\sigma$ , et le facteur de normalisation est  $n$  plutôt que  $n^{1/2}$ . Pour expliquer la présence de ces deux caractéristiques, considérons le cas simple, à savoir (20.06). Dans ce cas,

$$\hat{\alpha} = \frac{\sum y_t y_{t-1}}{\sum y_{t-1}^2},$$

où la somme s'applique aux observations allant de 1 à  $n$  à condition que  $y_0$  soit disponible, et de 2 à  $n$  dans le cas contraire. Nous supposons que  $y_0$  est disponible, puisque cela simplifie quelques résultats, et nous supposons également que les données sont générées par la marche aléatoire

$$y_t = y_{t-1} + u_t, \quad u_t \sim \text{IID}(0, \sigma^2).$$

<sup>1</sup> La notation utilisée pour ces statistiques varie d'un auteur à l'autre. Nous préférons celle-ci parce qu'elle repose sur un mécanisme mnémotechnique:  $nc$  indique "sans constante,"  $c$  "avec constante," et  $ct$  "constante et tendance."

Cela implique que le DGP est en réalité un cas particulier du modèle estimé. Afin d'éviter une dépendance infinie vis-à-vis du passé, il est nécessaire de supposer que  $y_{-j}$  est égale à une certaine valeur pour un  $j \geq 0$  quelconque. Pour rester concrets et simples, nous supposons que  $y_{-1} = 0$ .

Sous ces hypothèses,

$$\hat{\alpha} = \frac{\sum y_{t-1}^2}{\sum y_{t-1}^2} + \frac{\sum u_t y_{t-1}}{\sum y_{t-1}^2} = 1 + \frac{\sum u_t y_{t-1}}{\sum y_{t-1}^2}.$$

En ordonnant les termes autrement, nous avons

$$\hat{\alpha} - 1 = \frac{\sum u_t y_{t-1}}{\sum y_{t-1}^2}. \quad (20.09)$$

Il est clair qu'à la fois  $u_t$  et  $y_{t-1}$  doivent être proportionnels à  $\sigma$ . Ainsi le numérateur et le dénominateur de (20.09) doivent être proportionnels à  $\sigma^2$ . Ces facteurs de proportionnalité s'éliminant, nous obtenons une distribution de  $\hat{\alpha} - 1$  indépendante de  $\sigma$ . Ce résultat repose sur l'hypothèse selon laquelle  $y_{-1}$  est nulle. Si  $y_{-1}$  prend une valeur non nulle, ce résultat n'est vrai qu'asymptotiquement.

La seconde caractéristique étrange de (20.08), à savoir que le facteur de normalisation est  $n$  plutôt que  $n^{1/2}$ , est quelque peu plus délicate à expliquer. Définissons tout d'abord le **processus de somme partielle**  $S_t$  comme

$$S_t = \sum_{s=0}^t u_s,$$

ce qui nous permet d'écrire<sup>2</sup>

$$y_t = y_{-1} + S_t = S_t.$$

En substituant  $S_{t-1}$  à  $y_{t-1}$  dans le membre de droite de (20.09), nous obtenons

$$\hat{\alpha} - 1 = \frac{\sum u_t S_{t-1}}{\sum S_{t-1} S_{t-1}}. \quad (20.10)$$

On peut écrire le numérateur de cette expression comme

$$\sum_{t=1}^n \left( \sum_{s=0}^{t-1} u_s u_t \right).$$

<sup>2</sup> Sans l'hypothèse de nullité de  $y_{-1}$ , la seconde égalité ne serait pas exacte, et les expressions qui suivent seraient plus compliquées. Cependant, les termes impliquant  $y_{-1}$  ne seraient pas de la plus haute importance et n'affecteraient donc pas les résultats finals. Dans les modèles (20.05) et (20.07), aucune hypothèse sur  $y_{-1}$  n'est nécessaire, parce que l'ajout d'un terme constant dans la régression signifie que les moyennes de toutes les variables ont été éliminées.

La somme entre parenthèses possède  $t$  termes:  $u_0u_t, u_1u_t, u_2u_t$ , et ainsi de suite jusqu'à  $u_{t-1}u_t$ . La somme totale comprend donc  $\sum_{t=1}^n t = \frac{1}{2}n(n+1)$ , soit  $O(n^2)$  termes. Puisque nous avons supposé que les aléas ne sont pas auto-corrélés, chacun de ces termes doit avoir une espérance nulle. Sous l'hypothèse qu'un théorème de la limite centrale s'applique à leur somme, l'ordre de cette somme sera la racine carrée de  $n^2$ . Ainsi la somme est  $O(n)$ .

De façon tout à fait comparable, le dénominateur de (20.10) peut s'écrire comme

$$\sum_{t=1}^n \left( \sum_{r=0}^{t-1} \sum_{s=0}^{t-1} u_r u_s \right).$$

Chaque double somme à l'intérieur des parenthèses possède  $t^2$  termes. Parmi ceux-ci,  $t$  sont de la forme  $u_s^2$ , et les  $t^2 - t$  restants ont une espérance nulle. Ainsi chaque double somme sera  $O(t)$ , et donc aussi  $O(n)$ , et aura une espérance positive. La sommation de  $n$  de ces doubles sommes produit donc une quantité qui doit être  $O(n^2)$ . Ainsi nous voyons que le membre de droite de (20.10) est  $O(n)/O(n^2) = O(n^{-1})$ . Nous concluons par conséquent que le facteur de normalisation  $n$  dans (20.08) est précisément ce qui est nécessaire pour garantir que la statistique de test  $z$  soit  $O(1)$  sous l'hypothèse nulle.

L'analyse des régressions (20.07) ou (20.05) est encore plus compliquée que pour (20.06), mais la conclusion est identique:  $\hat{\alpha} - 1$  doit être normalisé par un facteur de  $n$  plutôt que par un facteur de  $n^{1/2}$ . Cela montre assez clairement que la théorie asymptotique standard ne s'applique pas aux statistiques  $\tau$  dont  $\hat{\alpha} - 1$  est le numérateur. Et la théorie asymptotique standard ne s'applique certainement pas davantage aux statistiques  $z$  elles-mêmes. En réalité, comme nous le verrons dans la section qui suit, les six statistiques de test dont nous avons discuté jusqu'à présent ont toutes des distributions asymptotiques différentes.

Il n'y a aucune raison de baser les tests de racine unitaire uniquement sur les régressions (20.05), (20.06), ou (20.07). En particulier, il est parfaitement valable d'ajouter d'autres régresseurs non stochastiques, tels que les variables muettes saisonnières, dans ces régressions. Il n'est pas pertinent d'ajouter des variables muettes à (20.06), puisqu'il n'y a pas de terme constant dans le modèle sur lequel elle se base. Cependant, c'est une stratégie pertinente pour (20.05) ou (20.07). Parce que les variables muettes saisonnières sont du même ordre que la constante, leur présence ne modifie pas asymptotiquement les distributions des statistiques de test.

Il est également envisageable d'ajouter des puissances de la tendance. Le modèle stationnaire autour d'une tendance (20.01) peut se généraliser en ajoutant  $t^2$  en tant que variable supplémentaire, impliquant donc que  $y_t$  est stationnaire autour d'une tendance quadratique. Identiquement, la marche aléatoire avec dérive (20.02) peut se généraliser en ajoutant une tendance temporelle linéaire, permettant à la dérive de varier dans le temps. Un modèle combiné qui emboîte les deux modèles peut s'écrire, après la reparamétrisation



classique, comme

$$\Delta y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + (\alpha - 1)y_{t-1} + u_t. \quad (20.11)$$

Comme on peut s'y attendre d'après ce qui survient pour (20.05) et (20.07),  $\beta_2 = 0$  lorsque  $\alpha = 1$  dans ce modèle. Les tests basés sur (20.11), et sur des équations possédant encore plus de puissances de la tendance, furent préconisés par Ouliaris, Park, et Phillips (1989). Les deux statistiques de test pour  $\alpha = 1$  basées sur (20.11) seront notées  $z_{ctt}$  et  $\tau_{ctt}$ , où *ctt* indique “avec constante, tendance, et tendance quadratique.” Parce que la tendance quadratique augmente plus rapidement avec  $t$  que ne le font la constante et la tendance linéaire, les distributions asymptotiques de ces tests sont différentes de celles des autres tests que nous avons abordés.

### 20.3 THÉORIE ASYMPTOTIQUE ET TESTS DE RACINE UNITAIRE

La théorie asymptotique pour des régressions qui impliquent des variables  $I(1)$ , ce qui comprend les régressions sur lesquelles sont basés les tests de racines unitaires, est très différente de la théorie asymptotique plus classique que nous avons exploitée jusqu'à présent dans cet ouvrage. Il est par conséquent impossible dans cette section de faire davantage qu'exposer quelques résultats importants et d'essayer de donner l'intuition de leur validité. Les articles de référence sont dans ce domaine ceux de Dickey et Fuller (1979), Phillips (1987), et Phillips et Perron (1988). Banerjee, Dolado, Galbraith, et Hendry (1993) apportent une introduction abordable des résultats de base.

Les théorèmes de la limite centrale classiques, tellement utiles pour les estimateurs qui approchent leur véritable valeur à des taux proportionnels à  $n^{-1/2}$ , ne sont plus d'aucune utilité avec les tests de racines unitaires. Au lieu de cela, il est nécessaire d'employer ce que l'on appelle des **théorèmes de la limite centrale fonctionnels**, parce qu'ils impliquent le calcul de la limite de certaines quantités dans un espace fonctionnel; voir Billingsley (1968) ou Hall et Heyde (1980). Nous n'essaierons pas de démontrer un quelconque théorème de la limite centrale fonctionnel, ni même de l'établir formellement. Cependant, nous tenterons de donner l'intuition de tels théorèmes dans ce contexte.

L'idée fondamentale qui permet l'utilisation des théorèmes de la limite centrale fonctionnels est l'idée d'une application d'une suite  $\{0, 1, 2, \dots, n\}$ , qui indice les observations, vers l'espace fermé  $[0, 1]$ . Supposons que l'on divise cet intervalle en  $n + 1$  portions, avec des divisions en  $1/(n + 1)$ ,  $2/(n + 1)$ , et ainsi de suite. Nous pouvons donc associer l'observation 0 au sous-intervalle  $0 \leq r < 1/(n + 1)$ , l'observation 1 au sous-intervalle  $1/(n + 1) \leq r < 2/(n + 1)$ , et ainsi de suite. Au fur et à mesure que  $n$  augmente et tend vers l'infini,

chaque sous-intervalle tend vers zéro. Ainsi si  $[rn]$  désigne l'entier le plus grand inférieur à  $rn$ , pour  $r \in [0, 1]$ , nous trouvons que

$$\begin{aligned} [r(n+1)] &= 0 \quad \text{pour } 0 \leq r < \frac{1}{n+1}, \\ [r(n+1)] &= 1 \quad \text{pour } \frac{1}{n+1} \leq r < \frac{2}{n+1}, \end{aligned}$$

et ainsi de suite jusqu'à

$$[r(n+1)] = n \quad \text{pour } \frac{n}{n+1} \leq r < 1.$$

Ainsi chaque réel  $r$  dans l'intervalle  $[0, 1]$  est associé à un et un seul indice  $0, 1, \dots, n$ .

Considérons à présent le **processus de somme partielle standardisé**

$$R_n(r) \equiv \frac{1}{\sigma\sqrt{n}} S_{[r(n+1)]} \equiv \frac{1}{\sigma\sqrt{n}} \sum_{s=0}^{[r(n+1)]} u_s, \quad r \in [0, 1].$$

Il s'agit simplement du processus de somme partielle ordinaire rencontré dans la section précédente, divisé par l'écart type des  $u_t$  et par la racine carrée de la taille de l'échantillon, et indicé par  $r$  plutôt que par  $t$ . On peut montrer à l'aide d'un théorème de la limite centrale fonctionnel que, sous des conditions relativement souples sur les  $u_t$ ,  $R_n(r)$  converge vers ce que l'on appelle un **processus de Wiener standard** et que l'on note  $W(r)$ . Intuitivement, un processus de Wiener est comparable à une marche aléatoire continue définie sur l'intervalle  $[0, 1]$ . Malgré sa continuité, il varie de façon erratique à chaque sous-intervalle, et chaque incrément est indépendant des autres. Une propriété quelquefois intéressante est que pour un  $r$  fixé,  $W(r) \sim N(0, r)$ .

Les principaux résultats sur les propriétés asymptotiques des statistiques de tests de racines unitaires sont que, sous l'hypothèse nulle de racine unitaire, elles convergent vers des fonctions variées des processus de Wiener. Malheureusement, de telles fonctions possèdent des distributions que l'on ne peut pas en général exprimer de manière commode, et doivent être évaluées numériquement. Pour donner une idée de l'aspect des résultats théoriques sur les propriétés asymptotiques des statistiques de test, nous énonçons les principaux résultats de Phillips (1987) pour les statistiques  $z_{nc}$  et  $\tau_{nc}$ :

$$z_{nc} \Rightarrow \frac{\frac{1}{2}(W^2(1) - 1)}{\int_0^1 W^2(r) dr} \quad (20.12)$$

$$\tau_{nc} \Rightarrow \frac{\frac{1}{2}(W^2(1) - 1)}{\left(\int_0^1 W^2(r) dr\right)^{1/2}}. \quad (20.13)$$

Ici le symbole  $\Rightarrow$  désigne la convergence faible dans un espace fonctionnel, qui est l'analogie de la convergence en distribution. Des résultats comparables pour les statistiques de test  $z_c$ ,  $z_{ct}$ ,  $\tau_c$ , et  $\tau_{ct}$  sont détaillés par Phillips et Perron (1988).

L'une des caractéristiques majeures de ces résultats est qu'ils ne dépendent pas de l'hypothèse d'homoscédasticité des aléas  $u_t$ . Les distributions asymptotiques des statistiques de test dont nous avons discuté sont identiques que les aléas manifestent une hétéroscédasticité de forme inconnue ou soient homoscédastiques. Malgré tout, il est essentiel qu'il n'y ait aucune corrélation entre  $u_t$  et  $u_{t-j}$  pour tout  $j \neq 0$ . Ainsi les statistiques de test dont nous avons parlé ne sont pas valables lorsque les aléas sont autocorrélés. En présence d'autocorrélation, il faut adapter les statistiques de test pour en tenir compte. Nous discuterons dans la section suivante de deux moyens de les modifier.

Bien que des résultats comme (20.12) et (20.13) soient d'un intérêt théorique considérable, ils ne sont pas très utiles dans la pratique, parce que les distributions des quantités du membre de droite ne sont pas connues analytiquement. Toutefois, des valeurs critiques pour les huit statistiques de test examinées ont été tabulées à l'aide de méthodes numériques nombreuses, dont les simulations Monte Carlo. La référence la plus connue est Fuller (1976), où quelques valeurs critiques asymptotiques pour  $\tau_{nc}$ ,  $\tau_c$ ,  $\tau_{ct}$ , ainsi que celles correspondant aux tests en  $z$ , sont tabulées, conjointement aux valeurs critiques en échantillon fini pour les quelques tailles d'échantillons retenues. Kiviet et Phillips (1990) montrent que les distributions en échantillon fini des tests en  $z$  peuvent se calculer numériquement, d'une manière très comparable à celle qui permet le calcul des distributions en échantillon fini de la statistique Durbin-Watson (Section 10.8), et ils tabulent quelques valeurs critiques à l'aide de cette méthode. Nabeya et Tanaka (1990) montrent comment on peut calculer analytiquement les distributions asymptotiques des statistiques  $z$  et tabulent un certain nombre de valeurs critiques pour  $z_{nc}$ ,  $z_c$ , et  $z_{ct}$ . MacKinnon (1991) emploie des méthodes Monte Carlo pour estimer des surfaces de réponse (voir la Section 21.7) pour quelques tests en  $\tau$ . Ces méthodes permettent une lecture immédiate des valeurs critiques pour n'importe quelle taille d'échantillon, aussi bien que pour  $n = \infty$ .

Hélas, toutes les valeurs critiques en échantillon fini pour les tests de racine unitaire dépendent d'au moins une hypothèse irréaliste sur les aléas, à savoir qu'ils sont NID( $0, \sigma^2$ ). Les valeurs critiques asymptotiques, au contraire, sont valables dans un contexte beaucoup plus général, puisqu'elles ne reposent ni sur la normalité ni sur l'homoscédasticité. Ainsi il semble plus prudent d'employer des valeurs critiques asymptotiques, de les traiter avec précaution, plutôt que de se fier à des valeurs d'échantillon fini qui peuvent se révéler tout à fait inadéquates dans la pratique.

Le Tableau 20.1 fournit quelques valeurs asymptotiques, calculées à l'aide de méthodes comparables à celles employées par MacKinnon (1991), pour les huit statistiques de test différentes abordées. La plupart des valeurs cri-

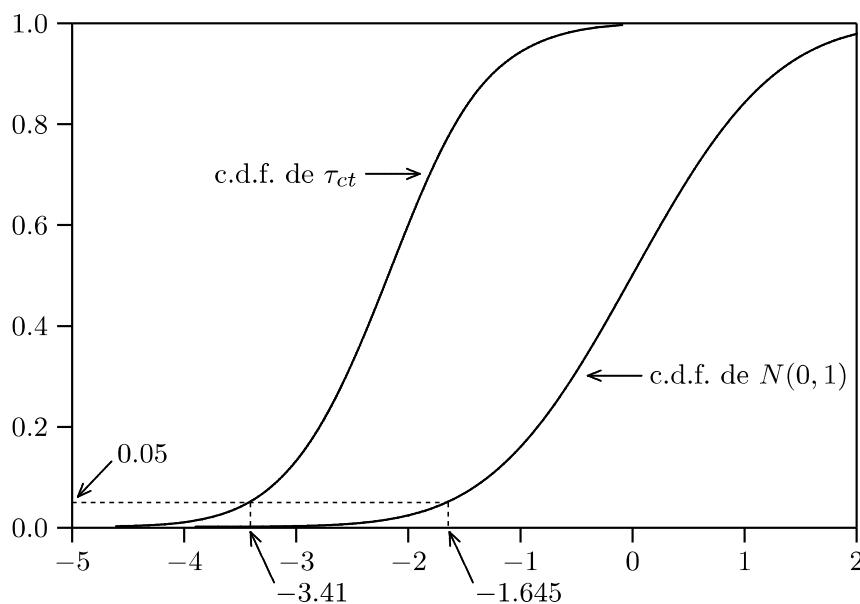
**Tableau 20.1** Valeurs Critiques Asymptotiques pour les Tests de Racine Unitaire

Statistique de Test	1%	2.5%	5%	10%	97.5%
$\tau_{nc}$	-2.56	-2.23	-1.94	-1.62	1.62
$\tau_c$	-3.43	-3.12	-2.86	-2.57	0.24
$\tau_{ct}$	-3.96	-3.66	-3.41	-3.13	-0.66
$\tau_{ctt}$	-4.37	-4.08	-3.83	-3.55	-1.21
$z_{nc}$	-13.7	-10.4	-8.0	-5.7	1.6
$z_c$	-20.6	-16.9	-14.1	-11.2	0.4
$z_{ct}$	-29.4	-25.1	-21.7	-18.2	-1.8
$z_{ctt}$	-36.6	-31.8	-28.1	-24.2	-4.2

tiques correspondent à celles de la queue de gauche de la distribution, étant donné que l'hypothèse alternative contre laquelle le test de racine unitaire est mené est presque toujours que le processus est stationnaire, plutôt qu'explosif. Ces valeurs diffèrent légèrement de celles publiées par Fuller (1976). Les différences, que l'on peut attribuer en première approximation à l'aspect aléatoire de la simulation, ne sont jamais supérieures à deux unités dans le dernier chiffre pertinent, et ne devraient donc pas avoir de conséquence dans les applications pratiques.

Il est clair d'après le Tableau 20.1 que le comportement asymptotique des statistiques de test de racine unitaire est très différent du comportement de n'importe quelle autre statistique de test rencontrée jusqu'à présent. Supposons que  $\alpha_0$  désigne la véritable valeur de  $\alpha$ . Dans le cas stationnaire, lorsque  $|\alpha_0| < 1$ , un  $t$  de Student pour  $\alpha = \alpha_0$  serait asymptotiquement distribué suivant la  $N(0, 1)$  sous l'hypothèse nulle. Ainsi les valeurs critiques à 2.5% et 97.5% pour un tel test seraient  $\pm 1.96$ . On peut comparer ces valeurs avec les valeurs critiques des tests en  $\tau$  données par le tableau. Les valeurs critiques à 2.5% sont toujours inférieures à  $-1.96$  et deviennent de plus en plus faibles lorsque l'on ajoute des régresseurs à la régression de test. Idem, les valeurs critiques à 97.5% sont toujours inférieures à  $1.96$  et en fait négatives pour les statistiques de test  $\tau_{ct}$  et  $\tau_{ctt}$ .

La Figure 20.1 illustre la fonction de répartition de la statistique  $\tau_{ct}$  pour le cas où  $n = 1000$ , qui est pratiquement indiscernable du cas asymptotique. Cette courbe trace en fait les points obtenus empiriquement par une expérience Monte Carlo; compte tenu du nombre de simulations, qui était de 5 millions, l'erreur expérimentale est négligeable. Par comparaison, nous avons également reporté la fonction de répartition de la normale centrée et réduite. Les différences entre les deux sont frappantes, la c.d.f. de  $\tau_{ct}$  basée sur une expérience Monte Carlo étant toujours bien à gauche de celle de la normale centrée réduite. La principale raison de ce décalage provient du fait que  $\hat{\alpha}$  est sérieusement biaisée vers 0 lorsque  $\alpha_0 = 1$ . Ce biais provoque des conséquences graves sur la puissance de ces tests à rejeter l'hypothèse nulle



**Figure 20.1** Distribution de  $\tau_{ct}$  pour  $n = 1000$

de racine unitaire. Par exemple, si l'on effectue un test unilatéral au niveau de 5%, les valeurs critiques asymptotiques pour  $z_c$ ,  $z_{ct}$ , et  $z_{ctt}$  sont, respectivement,  $-14.1$ ,  $-21.7$ , et  $-28.1$ . Ainsi, si  $n = 100$ ,  $\hat{\alpha}$  doit être inférieure à  $0.859$ ,  $0.783$ , et  $0.719$  dans ces trois situations pour que l'hypothèse nulle soit rejetée. A l'évidence, la puissance des tests de racine unitaire peut être faible si les données sont en réalité générées par un modèle stationnaire en tendance dont les aléas sont autocorrélés.

Nous avons noté à plusieurs reprises que, sous l'hypothèse nulle que  $\alpha = 1$ , les paramètres  $\beta_0$  dans la régression (20.07),  $\beta_1$  dans la régression (20.05), et  $\beta_2$  dans la régression (20.11) doivent être nuls. Notons  $\beta_k$  les paramètres qui doivent être nuls dans une régression de test; ici  $k = 0$  si seulement une constante est ajoutée, et  $k$  est égal au nombre des termes de tendance ajoutés dans le cas contraire. Le résultat que  $\beta_k = 0$  provient directement de la manipulation algébrique qui conduit à ces régressions en tant que versions reparamétrisées de régressions telles que (20.03), puisque  $\beta_k = (1 - \alpha)\gamma_k$ . Cependant, il existe une explication beaucoup plus profonde. La présence d'une racine unitaire accroît l'ordre de  $y_t$ . Il en va de même lorsque l'on ajoute une constante, une tendance, et une tendance quadratique. Si l'on veut préserver l'ordre de  $y_t$  dans l'hypothèse nulle de racine unitaire et dans l'hypothèse alternative de stationnarité autour d'une tendance, il est nécessaire d'ajouter à la régression de test un certain régresseur déterministe associé à un coefficient nul sous l'hypothèse nulle et non nul sous l'hypothèse alternative. Par exemple, considérons (20.05), pour laquelle  $k = 1$ . Sous l'hypothèse nulle, cette régression devient

$$\Delta y_t = \beta_0 + \beta_1 t + u_t.$$

Sous l'hypothèse alternative de stationnarité, nous savons que  $\Delta y_t$  doit être  $O(1)$ . Par ailleurs, le terme de tendance est  $O(n)$ . Le seul moyen de conserver l'ordre de  $\Delta y_t$  dans les hypothèses nulle et alternative est que  $\beta_1$  soit nul dans la première.

Tous les résultats asymptotiques des tests de Dickey-Fuller reposent sur l'hypothèse de nullité de  $\beta_k$ . Cette hypothèse peut être inadaptée lorsqu'il y a une racine unitaire uniquement lorsque le DGP *n'est pas* un cas particulier du modèle que l'on teste. Par exemple, si  $k = 0$  et si le DGP comprend un terme de dérive  $\gamma_1$ , la constante  $\beta_0$  dans le modèle que l'on teste serait non nulle. Dans tout cas comparable où  $\beta_k \neq 0$ , les résultats asymptotiques sont considérablement modifiés, comme l'a montré West (1988). En l'occurrence, dans de telles circonstances, les  $t$  de Student pour  $\alpha = 1$  sont véritablement distribués asymptotiquement suivant une normale centrée réduite.

Malgré la puissance de ce résultat, il n'est pas très utile. Il pose deux problèmes. En premier lieu, la distribution normale n'offre une bonne approximation aux distributions en échantillon fini des tests de racine unitaire en  $\tau$  que si  $\beta_k$  est important par rapport à  $\sigma$ . Hylleberg et Mizon (1989) et Kwiatkowski et Schmidt (1990) mettent ce résultat en évidence à l'aide d'expériences Monte Carlo dans les cas où  $k = 0$  et  $k = 1$ , respectivement. Lorsque  $\beta_k/\sigma$  et  $n$  sont dans l'ordre de grandeur que l'on rencontre habituellement dans les séries économiques chronologiques, ils trouvent que les distributions DF approximent beaucoup mieux les distributions des statistiques  $\tau$  que ne le fait la distribution normale centrée réduite. En second lieu, les tests de racine unitaire basés sur des régressions où  $\beta_k \neq 0$  manquent chroniquement de puissance. En vérité, pour  $k \geq 1$  la puissance de tels tests s'annule lorsque  $n \rightarrow \infty$ . Ainsi, asymptotiquement, ils ne rejettent jamais l'hypothèse nulle lorsqu'elle est inexacte, bien qu'ils puissent la rejeter lorsqu'elle est vraie. Perron (1988) et Campbell et Perron (1991) discutent de ce résultat.

## 20.4 AUTOCORRÉLATION ET PROBLÈMES CONNEXES

Tous les tests de racine unitaire rencontrés jusqu'à présent ne sont valables que sous l'hypothèse de non autocorrélation des aléas des régressions de test. Cette hypothèse est très souvent peu pertinente, parce que les fonctions de régression dans les régressions de test ne dépendent d'aucune variable économique. Cela rend très probable une autocorrélation des aléas. Par conséquent, nous avons besoin de tests de racine unitaire qui sont valables (asymptotiquement) en présence d'autocorrélation. Il y a deux manières différentes de calculer de tels tests. Il se trouve, et cela peut paraître surprenant, que les nouveaux tests ont la même distribution asymptotique que certains des tests dont nous avons déjà discuté.

Les tests de racine unitaire les plus simples valables en présence d'autocorrélation de forme inconnue sont des versions modifiées des tests en  $\tau$  de

Dickey-Fuller. On les appelle souvent **tests de Dickey-Fuller augmentés**, ou **tests ADF**. Ils furent proposés initialement par Dickey et Fuller (1979) sous l'hypothèse que les aléas suivent un processus AR d'ordre inconnu. Un travail ultérieur de Said et Dickey (1984) et Phillips et Perron (1988) montra qu'ils sont valables asymptotiquement sous des conditions moins contraignantes. Considérons les régressions de test (20.05), (20.06), (20.07), ou (20.11). Nous pouvons écrire n'importe quelle régression sous la forme

$$\Delta y_t = \mathbf{X}_t \boldsymbol{\beta} + (\alpha - 1)y_{t-1} + u_t, \quad (20.14)$$

où  $\mathbf{X}_t$  est composée de l'ensemble des régresseurs non stochastiques correspondant à la régression de test: l'ensemble vide pour (20.06), une constante pour (20.07), une constante et une tendance linéaire pour (20.05), et ainsi de suite.

Supposons à présent, par souci de simplicité, que l'aléa  $u_t$  dans (20.14) obéisse au processus AR(1) stationnaire  $u_t = \rho u_{t-1} + \varepsilon_t$ . Alors (20.14) deviendrait

$$\begin{aligned} \Delta y_t &= \mathbf{X}_t \boldsymbol{\beta} - \rho \mathbf{X}_{t-1} \boldsymbol{\beta} + (\rho + \alpha - 1)y_{t-1} - \alpha \rho y_{t-2} + \varepsilon_t \\ &= \mathbf{X}_t \boldsymbol{\beta}^* + (\rho + \alpha - 1 - \alpha \rho)y_{t-1} + \alpha \rho(y_{t-1} - y_{t-2}) + \varepsilon_t \end{aligned} \quad (20.15)$$

$$= \mathbf{X}_t \boldsymbol{\beta}^* + (\alpha - 1)(1 - \rho)y_{t-1} + \alpha \rho \Delta y_{t-1} + \varepsilon_t. \quad (20.16)$$

Nous pouvons remplacer  $\mathbf{X}_t \boldsymbol{\beta} - \rho \mathbf{X}_{t-1} \boldsymbol{\beta}$  par  $\mathbf{X}_t \boldsymbol{\beta}^*$  dans (20.15), pour un choix quelconque de  $\boldsymbol{\beta}^*$ , parce que chaque colonne de  $\mathbf{X}_{t-1}$  appartient à  $\mathcal{S}(\mathbf{X})$ . Ceci provient du fait que  $\mathbf{X}_t$  ne peut comprendre que des variables déterministes telles que la constante, une tendance linéaire, et d'autres (voir la Section 10.9). Ainsi chaque composante de  $\boldsymbol{\beta}^*$  est une combinaison linéaire des composantes de  $\boldsymbol{\beta}$ .

L'équation (20.16) est une régression linéaire de  $\Delta y_t$  sur  $\mathbf{X}_t$ ,  $y_{t-1}$ , et  $\Delta y_{t-1}$ . C'est simplement la régression originelle (20.14), avec un régresseur supplémentaire,  $\Delta y_{t-1}$ . L'ajout de ce régresseur provoque le remplacement de l'aléa  $u_t$  autocorrélé par l'aléa  $\varepsilon_t$  non autocorrélé. La version ADF de la statistique  $\tau$ , que nous appellerons statistique  $\tau'$ , est simplement le  $t$  de Student ordinaire correspondant au test de nullité du coefficient de  $y_{t-1}$  dans (20.16). Si l'autocorrélation des aléas de (20.14) était modélisée complètement par un processus AR(1), la statistique  $\tau'$  aurait exactement la même distribution asymptotique que la statistique DF  $\tau$  ordinaire, pour une spécification de  $\mathbf{X}_t$  identique. Le fait que le coefficient de  $y_{t-1}$  soit  $(\alpha - 1)(1 - \rho)$  plutôt que  $\alpha - 1$  n'est pas un problème en soi. Parce que nous avons supposé que  $|\rho| < 1$ , ce coefficient ne peut être nul que si  $\alpha = 1$ . Ainsi un test de nullité du coefficient de  $y_{t-1}$  est équivalent à un test de  $\alpha = 1$ .

Il est évidemment très aisé de calculer les statistiques  $\tau'$  à l'aide de régressions comme (20.16), mais il est plus difficile de calculer les statistiques  $z'$  correspondantes. Si le coefficient de  $y_{t-1}$  était multiplié par  $n$ , le résultat

serait  $n(\hat{\alpha} - 1)(1 - \hat{\rho})$  plutôt que  $n(\hat{\alpha} - 1)$ . Cette statistique de test n'aurait clairement pas la même distribution asymptotique que  $z$ . Bien qu'il soit possible de calculer des statistiques  $z'$  à partir de régressions telles que (20.16), cela est loin d'être facile à réaliser; consulter Dickey, Bell, et Miller (1986). Ainsi, dans la pratique, les tests en  $\tau'$  sont plus largement répandus alors que les tests en  $z'$  ne sont presque jamais employés.

Dans cet exemple simple, nous pouvons gérer l'autocorrélation en ajoutant un régresseur,  $\Delta y_{t-1}$ , à la régression de test. Il est aisé de voir que si  $u_t$  obéit à un processus  $AR(p)$ , nous devrions associer  $p$  régresseurs supplémentaires à la régression,  $\Delta y_{t-1}$ ,  $\Delta y_{t-2}$ , et ainsi de suite jusqu'à  $\Delta y_{t-p}$ . Mais que se passe-t-il si les aléas suivent un processus MA ou ARMA? Dans ces cas, la composante de moyenne mobile des aléas ne serait modélisée que par un processus AR d'ordre infini, de sorte qu'il semble falloir ajouter une infinité de valeurs retardées de  $\Delta y_t$ . Cela est impossible, bien évidemment. Par chance, nous n'avons pas besoin de recourir à une procédure aussi radicale. Comme l'ont montré Said et Dickey (1984), on peut utiliser à raison les tests ADF même lorsqu'il y a une composante de moyenne mobile dans les aléas, à condition de laisser tendre le nombre des retards de  $\Delta y_t$  compris dans la régression vers l'infini à un taux inférieur à  $n^{1/3}$ . Il s'agit simplement de considérer que les aléas suivent un processus  $AR(p)$ , et de faire en sorte que la croissance de  $p$  ne soit pas supérieure à  $n^{1/3}$ .

Dans la pratique, bien sûr, étant donné que  $n$  est fixé et ne tend pas vers l'infini, la connaissance du taux critique de  $n^{1/3}$  n'aide pas beaucoup au choix de  $p$ . De plus, un économètre ne connaît pas le processus qui a réellement généré les aléas. Ainsi, la stratégie habituelle consiste à ajouter autant de retards de  $\Delta y_t$  qu'il est nécessaire pour éliminer une quelconque autocorrélation des aléas. Les expériences Monte Carlo (Schwert, 1989) suggèrent que les tests ADF réalisent de bonnes performances sous l'hypothèse nulle même lorsque le processus générateur des aléas comprend une composante de moyenne mobile.

Le second moyen d'obtenir des statistiques de test de racine unitaire variables malgré la présence d'une autocorrélation de forme inconnue réside dans l'emploi des **tests de racine unitaire non paramétriques** de Phillips (1987) et Phillips et Perron (1988). Dans cette approche, les statistiques de test sont basées sur la régression de test d'origine (20.14), mais elles sont modifiées de telle manière que l'autocorrélation ne perturbe pas leurs distributions asymptotiques. Ces tests sont dénommés "non paramétriques" parce qu'aucune spécification du processus générateur des aléas n'est nécessaire.

La statistique  $z$  non paramétrique correspondant à une spécification quelconque de la matrice  $\mathbf{X}$  dans (20.14) peut s'écrire

$$z^* = n(\hat{\alpha} - 1) - \frac{n^2(\hat{\omega}^2 - \hat{\sigma}^2)}{2\mathbf{y}^\top \mathbf{M}_X \mathbf{y}}. \quad (20.17)$$

Cette statistique est simplement la statistique  $z$  ordinaire, corrigée d'un terme qui tend vers zéro asymptotiquement en l'absence d'autocorrélation. Ici,  $\hat{\sigma}^2$



désigne n'importe quelle estimation convergente de  $\sigma^2$  et  $\hat{\omega}^2$  n'importe quelle estimation convergente de

$$\omega^2 \equiv \lim_{n \rightarrow \infty} \left( \frac{1}{n} E(S_n^2) \right).$$

Sans autocorrélation,  $\omega^2 = \sigma^2$  du fait que

$$E(S_n^2) = E\left(\sum_{s=1}^n \sum_{t=1}^n u_s u_t\right) = n\sigma^2.$$

Avec autocorrélation, cependant,  $\omega^2$  différera de  $\sigma^2$ , parce que  $E(u_s u_t) \neq 0$  pour au moins un  $t \neq s$  quelconque.

Le calcul de  $z^*$  telle qu'elle est définie par (20.17) n'est pas entièrement immédiat, parce qu'il y a un choix multiple pour  $\hat{\omega}^2$ . Le problème de l'estimation de  $\omega^2$  est identique à celui de l'estimation des matrices de covariance en présence d'hétéroscédasticité et d'autocorrélation de formes inconnues. Nous avons vu la manière de procéder dans la Section 17.5. Une technique particulièrement simple fût suggérée par Newey et West (1987a). Grâce à celle-ci, l'estimation de  $\omega^2$  est

$$\hat{\omega}^2 = \frac{1}{n} \left( \sum_{t=1}^n \hat{u}_t^2 + 2 \sum_{j=1}^p w_{jp} \left( \sum_{t=j+1}^n \hat{u}_t \hat{u}_{t-j} \right) \right), \quad (20.18)$$

où  $w_{jp} = 1 - j/(p+1)$ . D'autres fonctions de pondération pourraient convenir, tant qu'elles maintiennent la positivité de  $\hat{\omega}^2$ . Le paramètre  $p$  de troncature des retards ne doit pas croître à un taux supérieur à  $n^{1/4}$  afin que  $\hat{\omega}^2$  soit une estimation convergente de  $\omega^2$ .

Les statistiques  $\tau$  non paramétriques sont obtenues par une modification des statistiques  $\tau$  ordinaires identique à celle qui transforme  $z$  en  $z^*$ :

$$\tau^* = \frac{\hat{\sigma}\tau}{\hat{\omega}} - \frac{n(\hat{\omega}^2 - \hat{\sigma}^2)}{2\hat{\omega} \mathbf{y}^\top \mathbf{M}_X \mathbf{y}}. \quad (20.19)$$

Dès lors que les quantités nécessaires au calcul de  $z^*$  sont disponibles, il est aisé de calculer  $\tau^*$ . Cependant, quelques résultats empiriques — voir Phillips et Perron (1988) et Schwert (1989) — montrent que les statistiques  $z^*$  tendent à avoir plus de puissance que les statistiques ADF  $\tau'$  et  $\tau^*$  non paramétriques.

Puisque différents utilisateurs peuvent très bien choisir des valeurs différentes de  $p$ , ou employer des poids  $w_{jp}$  différents, ils peuvent obtenir des valeurs différentes de  $z^*$  ou  $\tau^*$  pour des données identiques. Ceci est tout à fait contrariant mais inévitable. Pour compliquer davantage les choses, il existe d'autres techniques d'estimation de  $\omega^2$ , en plus de celle que procure (20.18). Certaines d'entre elles possèdent de bonnes propriétés, mais d'autres

possèdent quelques défauts rédhibitoires; voir Andrews (1991a, 1991b) et Ouliaris, Park, et Phillips (1989), parmi d'autres auteurs. Les propriétés en échantillon fini de ces différentes techniques peuvent différer substantiellement. Toutefois, elles semblent être relativement pauvres pour au moins quelques spécifications du processus générateur des aléas (Schwert, 1989). Par ailleurs, les distributions asymptotiques des statistiques  $\tau'$  n'approximent pas toujours de façon satisfaisante leur comportement en échantillon fini, mais celui-ci n'est jamais aussi mauvais que le comportement des statistiques  $z^*$  et  $\tau^*$ .

Puisqu'il existe un grand nombre de façons de calculer des statistiques de test de racine unitaire non paramétriques, aucune ne possédant de bonnes propriétés en échantillon fini sous l'hypothèse nulle dans tous les cas, il est potentiellement dangereux de se fier à ces statistiques. Avant de procéder à des inférences importantes sur la base d'une ou de plusieurs d'entre elles, il serait judicieux de mener une expérience Monte Carlo (voir le Chapitre 21) pour évaluer leurs performances avec des données comparables à celles utilisées.

L'autocorrélation n'est pas le seul problème qui entrave le chemin de celui qui tente de calculer des statistiques de test de racine unitaire. Un problème extrêmement sérieux est que ces statistiques souffrent d'une incapacité quasi totale à rejeter l'hypothèse nulle lorsqu'elles sont employées sur des données désaisonnalisées à l'aide de filtres linéaires ou de méthodes propres aux agences de statistiques officielles. Dans la Section 19.6, nous discutons de la tendance des estimations OLS de  $\alpha$  dans la régression  $y_t = \beta_0 + \alpha y_{t-1} + u_t$  à être biaisées vers 1 lorsque  $y_t$  est une série désaisonnalisée. Ce biais est présent dans toutes les régressions de test rencontrées jusqu'ici. Même lorsque  $\hat{\alpha}$  n'est pas véritablement biaisée vers 1, elle le sera toujours plus que l'estimation correspondante employant des séries brutes. Etant donné que les distributions tabulées des statistiques de test se basent sur le comportement de  $\hat{\alpha}$  pour ce dernier cas de figure, il est fort probable que des statistiques de test calculées à l'aide de séries ajustées par saison rejeteront l'hypothèse nulle beaucoup moins souvent qu'elles ne le devraient, compte tenu des valeurs critiques du Tableau 20.1. C'est exactement ce que Ghysels et Perron (1992) trouvent après une série d'expériences Monte Carlo.

Si cela est possible, il faut éviter de manipuler des données ajustées par saison dans le calcul des tests de racine unitaire. Une possibilité consiste à employer des données annuelles. Cela peut provoquer un rétrécissement de l'échantillon, mais les conséquences de cette stratégie sont moins graves que ce que l'on peut craindre. Shiller et Perron (1985) insistent sur le fait que c'est davantage l'**étendue** des données (c'est-à-dire le nombre d'années couvert par l'échantillon) que le nombre des observations qui détermine la puissance des tests. La raison en est que si  $\alpha$  est en réalité positif, mais inférieur à 1, il sera plus proche de 1 lorsque les données sont observées plus fréquemment. Ainsi un test basé sur  $n$  observations annuelles peut n'avoir qu'un manque de puissance léger par rapport à un test basé sur  $4n$  observations brutes, et

même avoir un supplément de puissance par rapport à un test basé sur  $4n$  observations de données ajustées par saison.

Si l'on emploie des données mensuelles ou trimestrielles, il faudrait qu'elles ne fussent pas ajustées. Malheureusement, comme nous l'avons remarqué dans le Chapitre 19, des données brutes pour de nombreuses séries temporelles sont introuvables pour de nombreux pays. De plus, l'usage de variables non ajustées par saison rend pratiquement nécessaires l'emploi de variables muettes saisonnières dans la régression et la prise en compte d'une autocorrélation à l'ordre quatre ou douze.

Un second problème majeur avec les tests de racine unitaire est qu'ils sont sensibles à l'hypothèse de stabilité du processus générateur des données sur l'échantillon entier. Perron (1989) montra que la puissance des tests de racine unitaire chute brutalement si le niveau ou la tendance d'une série est modifié de manière exogène à un quelconque moment de la période d'observation. Bien que la série soit stationnaire sur les deux sous-échantillons, il est pratiquement impossible de rejeter l'hypothèse nulle qu'elle est  $I(1)$  dans de tels cas.

Perron proposa par conséquent des techniques que l'on peut employer pour tester les racines unitaires conditionnellement à des modifications exogènes en niveau ou en tendance. Ses tests s'effectuent en régressant  $y_t$  sur une constante, une tendance linéaire, et une ou deux variables muettes qui permettent soit à la constante soit à la tendance, soit aux deux, de varier à partir d'un point particulier du temps. Les résidus de ces régressions sont alors utilisés dans une régression comme (20.06), et les statistiques  $z$ ,  $\tau$ ,  $z^*$ , et  $\tau^*$  habituelles sont calculées. Les distributions asymptotiques de ces statistiques ne sont pas les mêmes que celles de  $z_{ct}$  et  $\tau_{ct}$ , contrairement à ce qu'elles seraient en l'absence de variables muettes dans les régressions initiales (à cause du Théorème FWL). Au lieu de cela, elles dépendent des variables muettes dont on se sert et de l'endroit où s'opère le changement dans l'échantillon. Des valeurs critiques asymptotiques sont tabulées par Perron (1989).

Un grand nombre de recherches empiriques, suite à l'article de Nelson et Plosser (1982), semble avoir montré que les racines unitaires caractérisent un grand nombre de séries macroéconomiques. Perron y opposa l'idée que la prise en compte de la grande dépression de 1929 (en ce qui concerne les séries annuelles antérieures à 1973) ou du choc pétrolier (en ce qui concerne les séries trimestrielles d'après-guerre) modifiait radicalement les résultats et montra que la plupart des séries macroéconomiques américaines ne possédaient pas de racine unitaire. Il n'est pas tout à fait évident que cette théorie polémique résiste à la multiplication des tests.

Il y a eu un développement important des travaux empiriques faisant appel aux tests de racine unitaire; les exemples majeurs sont Nelson et Plosser (1982), Mankiw et Shapiro (1985), Campbell et Mankiw (1987), Perron et Phillips (1987), et DeJong et Whiteman (1991). Du fait des nombreux problèmes dont nous avons discuté, et parce que des tests différents tendent à produire des résultats différents, il est difficile d'établir des inférences

définitives sur la présence ou l'absence de racines unitaires dans les séries économiques temporelles. Cela suggère que, lorsque l'on tente d'élaborer des modèles de régression que l'on estime à l'aide de séries temporelles possédant éventuellement une racine unitaire, il ne faudrait pas adopter une stratégie performante uniquement si les données sont soit  $I(0)$  soit  $I(1)$ . Nous reviendrons sur ce point dans la Section 20.8. Avant d'envisager ce problème, nous devons aborder le thème fondamental de la cointégration.

## 20.5 COINTÉGRATION

La théorie économique suggère souvent que certaines paires de variables économiques doivent être liées par une relation d'équilibre de long terme. Bien que ces variables puissent s'éloigner de l'équilibre un certain temps, on s'attend à ce que des forces économiques rétablissent en quelque sorte l'équilibre. On trouve parmi ces relations celle des taux d'intérêts aux actifs à échéances différentes, celle des prix de biens de consommation comparables dans des pays différents (si les taux de change sont stables en longue période), celle du revenu disponible et de la consommation, celle des dépenses gouvernementales et des impôts, celle des salaires et des prix, celle de la demande de monnaie et du niveau des prix, ou encore celle des prix spot et futur d'un bien. Il n'y a aucune raison de se limiter à des paires de variables, bien sûr, bien que cela soit plus facile à gérer. Il peut très bien exister des groupes de trois variables, ou quatre, ou même davantage, que l'on imagine liées par une relation de long terme.

La plupart des variables mentionnées dans le premier paragraphe sont  $I(1)$ , ou du moins donnent l'apparence d'être non stationnaires lorsque certains tests de racine unitaire (mais pas nécessairement tous) sont utilisés. Nous savons que des variables  $I(1)$  tendent à diverger lorsque  $n \rightarrow \infty$ , parce que leur variance non conditionnelle est proportionnelle à  $n$ . Ainsi il semble que de telles variables n'obéissent jamais à une quelconque relation d'équilibre de long terme. Cependant, il est possible que certaines variables soient  $I(1)$  et que, malgré cela, des combinaisons linéaires de ces variables soient  $I(0)$ . Si c'est le cas, on parle de variables **cointégrées**. Si deux ou plusieurs variables sont cointégrées, elles doivent suivre un sentier d'équilibre de long terme, bien qu'en court terme elles puissent diverger substantiellement de l'équilibre. Le concept de cointégration est fondamental à la compréhension des relations d'équilibre de long terme entre les variables économiques temporelles. C'est également un concept assez récent. La référence la plus lointaine est Granger (1981), l'article le plus connu étant Engle et Granger (1987), et deux articles relativement accessibles sont Hendry (1986) et Stock et Watson (1988a).

Supposons, par souci de simplicité, que nous nous intéressions à deux variables,  $y_{t1}$  et  $y_{t2}$ , chacune étant  $I(1)$ . Alors, dans le cas le plus simple,  $y_{t1}$  et  $y_{t2}$  seraient cointégrées s'il existait un vecteur  $\boldsymbol{\eta} \equiv [1 \quad -\eta_2]^\top$  tel que,

lorsque les deux variables sont en équilibre,

$$[\mathbf{y}_1 \quad \mathbf{y}_2]\boldsymbol{\eta} \equiv \mathbf{y}_1 - \eta_2 \mathbf{y}_2 = 0. \quad (20.20)$$

Ici  $\mathbf{y}_1$  et  $\mathbf{y}_2$  désignent les vecteurs de dimension  $n$  dont les éléments types sont respectivement  $y_{t1}$  et  $y_{t2}$ . Le vecteur de  $\boldsymbol{\eta}$  de dimension 2 est appelé **vecteur cointégrant**. A l'évidence, il n'est pas unique, puisque nous pourrions le multiplier par n'importe quel scalaire non nul sans rien changer aux résultats de (20.20).

D'un point de vue plus réaliste, on s'attend à ce que  $y_{t1}$  et  $y_{t2}$  varient aussi bien systématiquement qu'aléatoirement dans le temps. Ainsi, on peut espérer trouver dans (20.20) une constante, et peut-être un ou plusieurs termes de tendance. Si nous posons  $\mathbf{Y} = [\mathbf{y}_1 \quad \mathbf{y}_2]$ , (20.20) peut prendre en compte cette éventualité sous la forme

$$\mathbf{Y}\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad (20.21)$$

où, comme dans (20.14),  $\mathbf{X}$  désigne une matrice déterministe qui peut contenir certains éléments. Si elle est non nulle, la première colonne sera une constante, la deuxième, si elle existe, sera une tendance linéaire, la troisième, si elle existe, sera une tendance quadratique, et ainsi de suite. Puisque  $\mathbf{Y}$  peut contenir plus de deux variables, (20.21) constitue en fait un moyen très général d'exprimer la relation de cointégration entre n'importe quel nombre de variables.

Evidemment, on ne peut pas s'attendre à ce qu'une égalité comme (20.20) ou (20.21) soit strictement satisfaite en n'importe quel instant  $t$  du temps. Nous pouvons donc définir une **erreur d'équilibre**  $\nu_t$  telle que

$$\nu_t = \mathbf{Y}_t \boldsymbol{\eta} - \mathbf{X}_t \boldsymbol{\beta}, \quad (20.22)$$

où  $\mathbf{Y}_t$  et  $\mathbf{X}_t$  désignent respectivement les lignes  $t$  de  $\mathbf{Y}$  et de  $\mathbf{X}$ . Dans le cas particulier de (20.20), cette erreur d'équilibre serait simplement  $y_{t1} - \eta_2 y_{t2}$ . Les  $m$  variables  $y_{t1}$  à  $y_{tm}$  sont dites cointégrées s'il existe un vecteur  $\boldsymbol{\eta}$  tel que  $\nu_t$  dans (20.22) soit  $I(0)$ .

Cette propriété est, à première vue, tout à fait remarquable. Ainsi, il peut ne pas être immédiatement évident que l'on puisse générer des variables  $I(1)$  mais cointégrées. Il est sans doute utile de considérer un exemple. Soit le modèle bivarié suivant:

$$\begin{aligned} \lambda_1 y_{t1} - y_{t2} &= u_{t1}, & (1 - \rho_1 L)u_{t1} &= \varepsilon_{t1}, \\ y_{t1} - \lambda_2 y_{t2} &= u_{t2}, & (1 - \rho_2 L)u_{t2} &= \varepsilon_{t2}, \end{aligned} \quad (20.23)$$

où  $y_{t1}$  et  $y_{t2}$  sont des variables aléatoires et  $\lambda_1$  et  $\lambda_2$  des paramètres, et

$$\begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega}).$$

Lorsqu'à la fois  $\rho_1$  et  $\rho_2$  sont inférieurs à 1,  $y_1$  et  $y_2$  seront à l'évidence  $I(0)$ . Lorsqu'à la fois  $\rho_1$  et  $\rho_2$  sont égaux à 1,  $y_1$  et  $y_2$  seront  $I(1)$ , et elles ne seront pas cointégrées. Cependant, si un  $\rho_i$  quelconque était égal à 1, l'autre étant inférieur à 1, les deux variables seraient  $I(1)$ , mais elles seraient cointégrées. Par exemple, supposons que  $\rho_2 < 1$  et que  $\rho_1 = 1$ . Alors, le vecteur cointégrant serait  $[1 \quad -\lambda_2]$ , et l'erreur d'équilibre serait

$$u_{t2} = y_{t1} - \lambda_2 y_{t2} = \varepsilon_{t2} + \rho_2 u_{t-1,2}.$$

Tant que  $\rho_2 < 1$ , cette erreur d'équilibre sera stationnaire et  $y_1$  et  $y_2$  seront cointégrées.

Le concept de cointégration porte en lui deux interrogations économétriques évidentes. La première concerne l'estimation du vecteur cointégrant  $\boldsymbol{\eta}$ , et la seconde concerne le test de deux ou plusieurs variables cointégrées. Ces questions sont bien sûr étroitement liées; la réponse à la seconde dépend de celle à la première. Nous verrons la première réponse dans les lignes qui suivent, et la seconde sera l'objet de la prochaine section.

Le moyen le plus simple d'estimer un vecteur cointégrant consiste à récrire (20.22) sous la forme d'une régression et à employer des OLS. Cette approche est associée à Engle et Granger (1987). Ainsi, si le coefficient de  $\mathbf{y}_1$  était arbitrairement normalisé à 1, nous pourrions exécuter la régression

$$\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \mathbf{Y}^*\boldsymbol{\eta}^* + \boldsymbol{\nu}, \quad (20.24)$$

où  $\mathbf{Y}^*$  est une matrice de dimension  $n \times (m-1)$  dont les colonnes sont  $\mathbf{y}_2$ ,  $\mathbf{y}_3$ , jusqu'à  $\mathbf{y}_m$ , et où le vecteur de paramètres  $\boldsymbol{\eta}^*$  est égal à l'opposé des  $m-1$  éléments non contraints du vecteur de paramètres  $\boldsymbol{\eta}$  qui apparaît dans (20.22).

Il y a en apparence deux problèmes majeurs dans l'exécution d'une régression comme (20.24). Le premier est que si les  $y_{it}$  sont cointégrées, elles sont sûrement déterminées conjointement, ce qui implique qu'il est très peu probable que les aléas soient indépendants des régresseurs. Dans le cas de (20.23), avec  $\rho_1 = 1$  et  $\rho_2 < 1$ , par exemple, la relation entre  $y_{t1}$  et  $y_{t2}$  est

$$y_{t1} = \lambda_2 y_{t2} + \rho_2 (y_{t-1,1} - \lambda_2 y_{t-1,2}) + \varepsilon_{t2}. \quad (20.25)$$

Ainsi, en régressant  $y_{t1}$  sur  $y_{t2}$ , le terme d'erreur est implicitement

$$\rho_2 (y_{t-1,1} - \lambda_2 y_{t-1,2}) + \varepsilon_{t2}, \quad (20.26)$$

et les deux termes sont corrélés à  $y_{t2}$ . Le second problème est que, dans une régression comme (20.24) nous régressons une variable  $I(1)$  sur une ou plusieurs autres variables  $I(1)$ . Cela semble être une stratégie peu recommandée, puisque c'est typiquement une situation où l'on rencontre des régressions erronées (voir la Section 19.2).

En dépit de ces deux problèmes, les estimations OLS de la régression (20.24) seront convergentes lorsque les variables  $y_{t1}$  à  $y_{tm}$  sont véritablement cointégrées. En fait, ces estimations seront **super-convergentes**; au lieu de converger vers la véritable valeur à un taux proportionnel à  $n^{-1/2}$ , elles convergeront à un taux proportionnel à  $n^{-1}$ . Le premier problème n'a pas d'importance asymptotiquement, puisque  $y_{t2}$  est  $I(1)$  et que les deux composantes du terme d'erreur dans (20.26) sont  $I(0)$  (la première composante n'est  $I(0)$  que si  $y_{t1}$  et  $y_{t2}$  sont véritablement cointégrées). Par conséquent les termes qui comprennent des aléas seront asymptotiquement négligeables relativement aux termes qui comprennent  $y_{t2}$ . Le second problème apparent ne se pose pas asymptotiquement pour des raisons comparables, à savoir que la (véritable) relation de cointégration entre les variables  $y_{ti}$  génère des termes qui dominent tout terme pouvant provoquer d'ordinaire une régression erronée. Une autre conséquence de tout ceci est que le  $R^2$  de (20.24) tendra vers 1 lorsque  $n \rightarrow \infty$ .

Pour comprendre la super-convergence des estimations de la régression (20.24), considérons le cas le plus simple, où  $m = 2$  et  $\mathbf{X}$  est une matrice nulle. Dans cette configuration, l'estimation OLS de  $\eta_2$ , le seul élément de  $\boldsymbol{\eta}^*$ , sera

$$\hat{\eta}_2 = \frac{\sum_{t=1}^n y_{t1} y_{t2}}{\sum_{t=1}^n y_{t2}^2}.$$

Si les deux séries sont cointégrées, nous avons

$$y_{t1} = \eta_2 y_{t2} + \nu_t,$$

où les  $\nu_t$  obéissent à un processus stationnaire quelconque. Par conséquent,

$$\hat{\eta}_2 = \eta_2 + \frac{\sum_{t=1}^n \nu_t y_{t2}}{\sum_{t=1}^n y_{t2}^2}. \quad (20.27)$$

Puisque  $y_{t2}$  est  $I(1)$ , nous l'exprimons comme

$$y_{t2} = S_{t2} + v_{t2},$$

où  $S_{t2}$  est un processus de somme partielle et où  $v_{t2}$  est une erreur qui serait i.i.d. si  $y_{t2}$  était une marche aléatoire, mais qui sera en général autocorrélée. Ainsi le second terme dans (20.27) est

$$\frac{\sum_{t=1}^n (\nu_t v_{t2} + \nu_t S_{t2})}{\sum_{t=1}^n (S_{t2}^2 + 2S_{t2}v_{t2} + v_{t2}^2)}. \quad (20.28)$$

On peut montrer, par des arguments similaires à ceux invoqués dans la Section 20.2, que les deux termes du numérateur sont  $O(n)$ . Le terme d'ordre

dominant dans le dénominateur est le premier, qui est  $O(n^2)$ . Ainsi, le rapport (20.28) est  $O(n)/O(n^2) = O(n^{-1})$ . Cela nous permet de conclure que  $\hat{\eta}_2$  converge vers la véritable valeur de  $\eta_2$  à un taux proportionnel à  $n^{-1}$ .

Ce résultat est crucial, et il se généralise au cas où  $\boldsymbol{\eta}$  est un vecteur à  $m$  composantes; voir Stock (1987). Il existe  $m$  manières d'exécuter une régression comme (20.24), correspondant chacune au  $\mathbf{y}_i$  que l'on place en régressande. Cela produira  $m$  vecteurs cointégrants estimés différents, tous étant super-convergens. Etant donné que des régressions ne comprenant que des séries stationnaires produisent toujours des estimations convergentes au taux  $n^{-1/2}$ , il est toujours possible de remplacer  $\boldsymbol{\eta}$  par  $\hat{\boldsymbol{\eta}}$  dans de telles régressions sans perturber leurs propriétés asymptotiques. Parce que les différences entre  $\boldsymbol{\eta}$  et  $\hat{\boldsymbol{\eta}}$  seront  $O(n^{-1})$ , nous pouvons les négliger asymptotiquement face aux erreurs d'estimations de telles régressions.

Malheureusement, la super-convergence de  $\hat{\boldsymbol{\eta}}$  *n'implique pas* qu'il possède toujours de bonnes propriétés en échantillon fini. Une partie du problème provient du fait que l'expression (20.28) n'a pas une espérance nulle, ce qui provoquera, en général, un biais de  $\hat{\boldsymbol{\eta}}$ . Ce biais peut être important dans la pratique; consulter Banerjee, Dolado, Hendry, et Smith (1986) et Stock (1987). Une source de biais est évidente si l'on examine (20.25). Cette équation comprend le terme  $\rho_2(y_{t-1,1} - \lambda_2 y_{t-1,2})$ , dont nous ne tenons pas compte en régressant  $y_{t1}$  sur  $y_{t2}$ . Le terme omis ressemble à un terme de correction d'erreur. Puisqu'il est  $I(0)$  et que  $y_{t2}$  est  $I(1)$ , sa mise à l'écart n'a que peu d'importance asymptotiquement. Par contre, lorsque  $\rho_2$  est important, il peut y avoir une corrélation importante entre  $y_{t-1,1} - \lambda y_{t-1,2}$  et  $y_{t1}$  en échantillon fini. Dans ce cas, cela peut provoquer un biais et une perte d'efficacité.

Des procédures d'amélioration des estimations de  $\boldsymbol{\eta}$  furent proposées par de nombreux auteurs, dont Phillips et Hansen (1990) et Saikkonen (1991). L'approche de ce dernier est particulièrement élégante. Il démontre que l'on peut obtenir des estimations asymptotiquement efficaces en exécutant la régression

$$\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \mathbf{Y}^*\boldsymbol{\eta}^* + \sum_{j=-p}^p \Delta \mathbf{Y}_{-j}^* \boldsymbol{\gamma}_j + \mathbf{e} \quad (20.29)$$

par moindres carrés. Ici,  $\Delta \mathbf{Y}_{-j}^*$  désigne une matrice de dimension  $n \times (m-1)$ , dont chaque colonne est un vecteur de différences premières de la colonne correspondante dans  $\mathbf{Y}^*$ , retardé de  $j$  périodes, et  $\boldsymbol{\gamma}_j$  désigne un vecteur composé de  $(m-1)$  coefficients. L'équation (20.29) ajoute simplement  $p$  avances et  $p$  retards des différences premières de  $\mathbf{Y}^*$  à la régression (20.24). Cette technique élimine les effets néfastes de la dynamique de courte période que les erreurs d'équilibre  $\boldsymbol{\nu}$  font subir aux estimations de  $\boldsymbol{\eta}$ . Parce que ces dernières ne sont pas asymptotiquement normalement distribuées, le concept d'efficacité employé par Saikkonen n'est pas le concept standard dont nous avons parlé dans cet ouvrage, et son article est loin d'être élémentaire. Bien sûr, son résultat



n'a de valeur qu'asymptotiquement. Si  $n$  n'est pas grand face à  $p(m-1)$ , il peut y avoir tellement de régresseurs supplémentaires dans (20.29) que les propriétés en échantillon fini des estimations par moindres carrés de  $\boldsymbol{\eta}^*$  peuvent être très médiocres.

## 20.6 TESTS DE COINTÉGRATION

Les tests de cointégration les plus familiers, qui sont étroitement reliés aux tests de racine unitaire, furent proposés par Engle et Granger (1987). L'idée de base est extrêmement simple. Si les variables  $y_{t1}$  à  $y_{tm}$  sont véritablement cointégrées, la véritable erreur d'équilibre  $\nu_t$  doit être  $I(0)$ . Si elles ne sont pas cointégrées, cependant,  $\nu_t$  doit être  $I(1)$ . Ainsi il est possible de tester l'hypothèse nulle de non existence d'une relation de cointégration contre l'hypothèse alternative de cointégration en exécutant un test de racine unitaire sur  $\nu_t$ .

Si  $\nu_t$  était observé, les tests de racine unitaire auraient la même distribution que ceux examinés précédemment. Toutefois, dans la grande majorité des cas, nous n'observerons pas  $\nu_t$  parce qu'au moins un élément de  $\boldsymbol{\eta}$  sera inconnu. Il est donc nécessaire d'estimer  $\boldsymbol{\eta}$ . Cela peut se faire en principe de plusieurs manières, la plus simple étant d'appliquer les OLS à la régression (20.24). Cette procédure fournit un vecteur de résidus, ou d'erreurs d'équilibre estimées,  $\hat{\nu}$ . Si les variables  $y_{t1}$  à  $y_{tm}$  sont en réalité non cointégrées, la régression (20.24) est falsifiée, et la série  $\hat{\nu}$  possède une racine unitaire. Les statistiques de test de racine unitaire classiques peuvent se calculer à l'aide du vecteur de résidus. Pour des raisons évidentes, ces tests sont appelés **tests de cointégration sur résidus**. Parce que  $\hat{\nu}$  dépend d'un ou de plusieurs paramètres estimés, qui sont les paramètres d'une régression falsifiée sous l'hypothèse nulle, les distributions asymptotiques des statistiques de test de cointégration sur résidus *ne sont pas* les mêmes que celles des statistiques de test de racine unitaire ordinaires.

Le modèle (20.23) peut procurer un éclaircissement utile. Puisque c'est la valeur de  $\rho_2$  (ou éventuellement celle de  $\rho_1$ ) qui détermine la cointégration entre les deux séries dans ce modèle, il ne devrait pas être surprenant d'apprendre que les tests de l'hypothèse nulle de non cointégration devraient ressembler aux tests de l'hypothèse nulle qu'une série possède une racine unitaire. Il ne devrait pas être surprenant non plus d'apprendre que l'hypothèse nulle est que les deux séries *ne sont pas* cointégrées, puisque, conditionnellement à  $\rho_1 = 1$ , elles seront cointégrées à moins que  $\rho_2$  ne soit égal à 1.

On peut adapter des tests de cointégration sur résidus à partir de n'importe lequel des tests de racine unitaire dont nous avons parlé, à condition toutefois d'employer des valeurs critiques appropriées. La procédure la plus simple, appelée parfois **test de Engle-Granger**, ou **test EG**, implique une première estimation de la régression de cointégration (20.24) et par la suite

l'usage d'un test de Dickey-Fuller en  $\tau$ , basé sur la régression

$$\Delta\hat{\nu}_t = (\alpha - 1)\hat{\nu}_{t-1} + e_t. \quad (20.30)$$

Puisque l'autocorrélation est très souvent un problème, on préférera employer un **test de Engle-Granger augmenté**, ou **test AEG**, qui est au test EG ce que le test ADF en  $\tau'$  est au test DF en  $\tau$ . Ainsi le test AEG est simplement le  $t$  de Student de  $\alpha - 1$  dans une régression comme (20.30) mais avec suffisamment de retards de  $\Delta\hat{\nu}_t$  comme régresseurs additionnels pour que toute autocorrélation soit éliminée. Des tests en  $z^*$  et en  $\tau^*$  non paramétriques peuvent également être utilisés, ainsi que l'on suggéré Phillips et Ouliaris (1990). Ceux-ci sont calculés exactement de la même manière que dans les expressions (20.17) et (20.19): les résidus de la régression (20.30) sont employés pour évaluer  $\hat{\sigma}^2$  et  $\hat{\omega}^2$ , et la quantité  $\hat{\nu}^\top \hat{\nu}$  remplace  $\mathbf{y}^\top \mathbf{M}_X \mathbf{y}$ .

Les valeurs critiques de ces tests dépendent du nombre de variables  $I(1)$  présentes dans le membre de droite de la régression de cointégration (20.24) ainsi que de la nature des régresseurs aléatoires dans cette régression. Quelques valeurs critiques relativement peu précises furent publiées par Engle et Granger (1987), Engle et Yoo (1987), et Phillips et Ouliaris (1990). Le Tableau 20.2 contient des valeurs critiques asymptotiques assez précises (la probabilité que l'erreur sur la dernière décimale soit supérieure à 2 est extrêmement faible) pour les statistiques  $\tau_c$ ,  $\tau_{ct}$ ,  $\tau_{ctt}$ ,  $z_c$ ,  $z_{ct}$ , et  $z_{ctt}$ , pour quelques valeurs de  $m$ , obtenues par des méthodes similaires à celle employée par MacKinnon (1991). Le tableau ne contient pas de valeurs critiques pour les statistiques  $\tau_{nc}$  ou  $z_{nc}$ , parce que cela est rarement pertinent dans la pratique. Souvenons-nous que  $m$  est le nombre de variables endogènes;  $m - 1$  est par conséquent le nombre d'éléments du vecteur cointégrant qu'il s'agit d'estimer. Si certains éléments sont connus a priori, il faut sélectionner une valeur de  $m$  plus faible. Dans le cas extrême où tous les éléments du vecteur cointégrant sont connus, il faudrait se reporter aux valeurs critiques du Tableau 20.1.

Parce que les régressions de cointégration contiennent les colonnes de  $\mathbf{X}$  parmi les régresseurs, il n'est pas nécessaire d'inclure  $\mathbf{X}$  dans la régression de test (20.30). Le Théorème FWL ne s'applique pas ici, parce que l'élimination de la première observation signifie que le vecteur  $\hat{\nu}_{-1}$  ne sera pas véritablement orthogonal aux colonnes de  $\mathbf{X}$ . Cependant,  $\hat{\nu}_{-1}$  sera orthogonal à  $\mathbf{X}$ , asymptotiquement. Ainsi, asymptotiquement, que  $\mathbf{X}$  soit incluse ou non dans la régression n'a pas d'importance.

Les estimations OLS  $\boldsymbol{\eta}$  dépendent du  $\mathbf{y}_i$  qui est régressande. Un changement de régressande modifiera, avec des échantillons finis, le vecteur de résidus  $\hat{\nu}$  et par conséquent les valeurs calculées des statistiques de test de cointégration basées sur ce vecteur. Cela est plutôt gênant, parce que cela s'ajoute à la multiplicité des statistiques de test. Ainsi en ce qui concerne les tests de cointégration, plus encore qu'en ce qui concerne les tests de racine unitaire, les occasions de commettre des inférences divergentes sont nombreuses.

**Tableau 20.2** Valeurs Critiques Asymptotiques pour les Tests de Cointégration

Statistique de Test	1%	2.5%	5%	10%	97.5%
$m = 2$					
$\tau_c$	-3.90	-3.59	-3.34	-3.04	-0.30
$\tau_{ct}$	-4.32	-4.03	-3.78	-3.50	-1.03
$\tau_{ctt}$	-4.69	-4.40	-4.15	-3.87	-1.52
$z_c$	-28.3	-23.9	-20.6	-17.1	-0.7
$z_{ct}$	-35.8	-31.1	-27.3	-23.4	-3.2
$z_{ctt}$	-42.6	-37.5	-33.4	-29.1	-5.8
$m = 3$					
$\tau_c$	-4.29	-4.00	-3.74	-3.45	-0.85
$\tau_{ct}$	-4.66	-4.37	-4.12	-3.84	-1.39
$\tau_{ctt}$	-4.99	-4.70	-4.45	-4.17	-1.81
$z_c$	-35.2	-30.4	-26.7	-22.7	-2.4
$z_{ct}$	-42.0	-36.9	-32.8	-28.5	-5.0
$z_{ctt}$	-48.5	-43.0	-38.7	-34.0	-7.6
$m = 4$					
$\tau_c$	-4.64	-4.35	-4.10	-3.81	-1.30
$\tau_{ct}$	-4.97	-4.68	-4.43	-4.15	-1.73
$\tau_{ctt}$	-5.27	-4.98	-4.73	-4.45	-2.09
$z_c$	-41.6	-36.5	-32.4	-28.1	-4.5
$z_{ct}$	-48.1	-42.6	-38.2	-33.5	-7.0
$z_{ctt}$	-54.3	-48.5	-43.9	-38.9	-9.8
$m = 5$					
$\tau_c$	-4.96	-4.66	-4.42	-4.13	-1.68
$\tau_{ct}$	-5.25	-4.96	-4.72	-4.43	-2.04
$\tau_{ctt}$	-5.53	-5.24	-4.99	-4.72	-2.36
$z_c$	-47.8	-42.3	-38.0	-33.3	-6.7
$z_{ct}$	-54.0	-48.2	-43.5	-38.5	-9.3
$z_{ctt}$	-60.0	-53.9	-49.0	-43.7	-12.1
$m = 6$					
$\tau_c$	-5.25	-4.96	-4.71	-4.42	-2.01
$\tau_{ct}$	-5.52	-5.23	-4.98	-4.70	-2.32
$\tau_{ctt}$	-5.77	-5.49	-5.24	-4.96	-2.61
$z_c$	-53.8	-48.0	-43.4	-38.4	-9.1
$z_{ct}$	-59.7	-53.7	-48.8	-43.5	-11.8
$z_{ctt}$	-65.5	-59.2	-54.1	-48.6	-14.6

Tous les problèmes qui enveniment les tests de racine unitaire enveniment également les tests de cointégration sur résidus dont nous avons parlé. Un problème vient du fait que les valeurs critiques asymptotiques peuvent se révéler sérieusement trompeuses avec des échantillons finis. Malheureusement, les valeurs critiques dépendent des caractéristiques intrinsèques du DGP, telles que la nature d'une quelconque hétéroscédasticité ou autocorrélation que l'on pourrait y rencontrer, qui sont en général inconnues dans la pra-

tique. Un autre problème, introduit dans la Section 20.4, est que les tests de cointégration manquent chroniquement de puissance lorsque l'on emploie des données désaisonnalisées ou lorsque le processus générateur d'une série quelconque varie dans le temps. Ainsi le non rejet de l'hypothèse nulle de non cointégration ne procure qu'un renseignement limité sur le fait que deux variables sont véritablement non cointégrées.

Bien que les tests basés sur le vecteur de résidus  $\hat{\mathbf{v}}$  soient de loin les plus répandus, de nombreux autres tests de cointégration furent proposés. On pourra par exemple consulter Stock et Watson (1988b), Phillips et Ouliaris (1990), Johansen (1988, 1991), et Johansen et Juselius (1990, 1992). L'approche de Johansen sera abordée dans la Section 20.8. Campbell et Perron (1991) font un exposé des nombreux tests, qui sont beaucoup plus difficiles à calculer que ceux reposant sur les résidus. En plus, chaque statistique de test semble posséder son propre ensemble de valeurs critiques.

## 20.7 MODÉLISATIONS AVEC DES VARIABLES COINTÉGRÉES

De nombreuses séries économiques sont, ou du moins paraissent être, intégrées d'ordre 1. A partir des résultats de la Section 19.2 sur les régressions erronées, et des résultats de ce chapitre, il est clair que régresser une variable  $I(1)$  en niveau sur une ou plusieurs variables  $I(1)$  en niveaux n'est généralement pas la meilleure stratégie à suivre. Au pire, nous "découvririons" une relation entièrement fausse. Au mieux, nous estimerions de façon convergente les éléments d'un vecteur cointégrant quelconque, mais nous ne pourrions pas appliquer la théorie asymptotique standard, et commettrions donc des inférences inexactes à propos des paramètres que nous aurions estimés. L'étude des méthodes de spécification et d'estimation des modèles pour des variables  $I(1)$  est un champ de recherche florissant et quelque peu controversé. La plupart du matériel théorique, tel que celui de Park et Phillips (1988, 1989) et Phillips (1991a), est techniquement trop lourd pour être traité dans cet ouvrage. Dans cette section, nous nous contenterons donc d'exposer des cas particuliers simples et quelques résultats relativement immédiats. Nous traiterons de l'estimation des autorégressions vectorielles impliquant des variables cointégrées dans la section qui suit.

L'approche classique pour gérer des variables cointégrées, en particulier dans la littérature des séries temporelles, a consisté à en calculer les différences premières autant de fois que nécessaire pour les rendre stationnaires. Cette approche a le mérite de la simplicité. Une fois toutes les séries transformées et stationnalisées, nous pouvons spécifier des modèles de régression dynamiques de manière conventionnelle, et leur appliquer des résultats asymptotiques standards. Le problème relatif à cette approche est que le calcul des différences élimine automatiquement l'opportunité d'estimer une quelconque relation entre les *niveaux* de la variable dépendante et ceux des variables indépendantes. Au contraire la cointégration implique qu'une telle relation existe, et, comme

les exemples du début de la Section 20.5 le suggèrent, sont d'un intérêt économique majeur. Le calcul des différences sur les données n'est donc pas une stratégie appropriée.

Une seconde approche consiste à estimer une sorte de modèle à correction d'erreur, ou ECM. Nous avons vu ce genre de modèle dans la Section 19.4, sous l'hypothèse que toutes les variables étaient stationnaires. Les modèles à correction d'erreur restent valables lorsque cette hypothèse n'est plus vérifiée. En réalité, ils sont particulièrement attrayants lorsque la variable dépendante est  $I(1)$ . Cependant, il faut rester prudent lors de l'estimation et de la réalisation d'inférences avec de tels modèles.

Un modèle ECM univarié simple mais largement adaptable, comparable à l'équation (19.30), peut s'écrire comme

$$\Delta y_t = \mathbf{z}_t \boldsymbol{\alpha} + \beta(y_{t-1} - \lambda x_{t-1}) + \gamma \Delta x_t + u_t, \quad u_t \sim \text{IID}(0, \sigma^2). \quad (20.31)$$

La variable dépendante est ici  $y_t$ , et la variable indépendante est véritablement  $x_t$ . Nous supposons que ces deux variables sont  $I(1)$  et cointégrées, ce qui implique que le terme de correction d'erreur  $\beta(y_{t-1} - \lambda x_{t-1})$  est  $I(0)$ . Le vecteur ligne  $\mathbf{z}_t$  comprend une constante, et toute autre variable indépendante, à condition qu'elle soit ou bien déterministe ou bien  $I(0)$ . Si la dynamique que procure (20.31) n'est pas satisfaisante, il est possible d'adapter le modèle en lui ajoutant davantage de retards de  $\Delta x_t$  et en augmentant le retard du terme de correction d'erreur.

Si  $\lambda$  était connu, l'estimation par moindres carrés de (20.31) se ferait sans difficulté. La régressande et les régresseurs seraient soit déterministes soit  $I(0)$ . Ainsi les estimations de  $\boldsymbol{\alpha}$ ,  $\beta$ , et  $\gamma$  seraient convergentes au taux  $n^{-1/2}$  et asymptotiquement normales, et leur matrice de covariance serait estimée de manière habituelle. Mais dans de nombreux cas,  $\lambda$  sera inconnu. Il y a alors plusieurs manières de procéder. La plus simple est la **méthode de Engle-Granger en deux étapes** proposée par Engle et Granger (1987). La première étape consiste à régresser  $y_t$  sur  $x_t$ , une constante, et une tendance linéaire si celle-ci apparaît dans  $\mathbf{z}_t$ . Comme nous l'avons vu, cela produira une estimation super-convergente de  $\lambda$ , disons  $\tilde{\lambda}$ . La seconde étape consiste à remplacer  $\lambda$  par  $\tilde{\lambda}$  dans (20.31) et à estimer par OLS cette équation transformée. En exploitant la propriété de super-convergence de  $\tilde{\lambda}$ , Engle et Granger montrent que les estimations des autres paramètres sont asymptotiquement identiques à celles obtenues connaissant  $\lambda$ .

Le mérite majeur de la procédure en deux étapes de Engle-Granger est incontestablement sa simplicité. Cependant, des simulations Monte Carlo ont largement montré qu'elle peut ne pas être fiable avec des échantillons finis; consulter Banerjee, Dolado, Hendry, et Smith (1986) et Banerjee, Dolado, Galbraith, et Hendry (1993). Le problème est que  $\tilde{\lambda}$  semble être bien souvent sévèrement biaisé. Ce biais se transmet alors aux autres paramètres estimés. Le problème s'avère moins grave lorsque le  $R^2$  de la régression de cointégration

est proche de 1, ce qui doit être le cas avec une taille d'échantillon assez importante. Ainsi, une valeur relativement faible du  $R^2$  de la régression de cointégration est un signal d'alarme de défaillance de la procédure.

La plus simple des procédures alternatives à la méthode en deux étapes de Engle-Granger consiste à estimer le modèle

$$\Delta y_t = \mathbf{z}_t \boldsymbol{\alpha} + \beta y_{t-1} + \delta x_{t-1} + \gamma \Delta x_t + u_t, \quad (20.32)$$

dans lequel le nouveau paramètre  $\delta$  est implicitement égal à  $-\beta\lambda$ . Cette régression est intrigante, puisque la variable dépendante est  $I(0)$  et les régresseurs sont  $I(1)$ . On devrait normalement s'attendre à ce que la théorie de la distribution asymptotique standard ne s'applique pas à certaines estimations ou à toutes. S'il est vrai que la théorie de la distribution asymptotique pour cette équation est non standard, les problèmes pratiques se révèlent moins graves que ce que l'on pourrait craindre.

Les résultats fondamentaux pour des régressions telles que (20.32) ont été démontrés par Sims, Stock, et Watson (1990). Ils envisagent les distributions asymptotiques des coefficients individuels dans une régression linéaire impliquant des variables  $I(1)$ . Ils montrèrent que si un paramètre  $\theta$  est associé à une variable  $I(0)$  de moyenne nulle, la quantité  $n^{1/2}(\hat{\theta} - \theta_0)$  sera asymptotiquement distribuée suivant une loi normale, avec l'écart type asymptotique habituel. Considérons (20.32) une nouvelle fois. Dans cette équation,  $\gamma$  est associé à une variable  $I(0)$ . À condition que  $\mathbf{z}_t$  contienne un terme constant, la condition de moyenne nulle est aisément remplie. De plus, comme (20.31) le montre clairement, on peut associer  $\beta$  à  $y_{t-1} - \lambda x_{t-1}$ , qui est  $I(0)$  du fait que  $x$  et  $y$  sont cointégrées. Si nous normalisons une nouvelle fois la régression de cointégration de sorte que  $x_{t-1}$  soit associée à un coefficient unitaire, nous voyons que l'on peut associer  $\delta$  à une variable qui est  $I(0)$ . Ainsi la théorie de la distribution asymptotique standard s'applique à tous les coefficients économiquement pertinents de (20.32).

Bien que l'on puisse pratiquer des inférences sur les coefficients *individuels* dans l'équation (20.32) de manière usuelle, il faut être prudent si l'on tente d'en faire davantage. Par exemple, un test de nullité jointe de  $\beta$  et de  $\delta$ , ou d'égalité à toute autre valeur, *n'aurait pas* la distribution asymptotique du  $\chi^2$  habituelle. Dans un ordre d'idée différent, on peut choisir de calculer  $\lambda$  comme  $-\tilde{\delta}/\tilde{\beta}$ , où  $\tilde{\beta}$  et  $\tilde{\delta}$  désignent les estimations par moindres carrés. Puisque  $\lambda$  n'est pas un coefficient associé à une variable  $I(0)$  de moyenne nulle, la théorie de la distribution asymptotique standard ne s'applique plus.

L'estimation directe de (20.31) par moindres carrés non linéaires est équivalente à l'estimation de l'équation (20.32) par OLS. Les valeurs ajustées des deux équations seront identiques, ainsi que les estimations des paramètres qu'elles ont en commun. Les résultats de Banerjee, Dolado, Hendry, et Smith (1986) suggèrent que ces estimations seront meilleures que celles obtenues par la méthode en deux étapes de Engle-Granger, mais cette conclusion fut remise

en cause par Engle et Yoo (1987, 1991). Il semblerait que les mérites respectifs des deux procédures d'estimation dépendent fortement des caractéristiques précises du DGP.

Les techniques d'estimation abordées dans cette section s'appliquent à une seule équation, et elles ne sont pas efficaces en général. Bien que la procédure en deux étapes soit toujours super-convergente pour  $\lambda$ , elle n'est pas asymptotiquement efficace. A la fin de la Section 20.5, nous avons introduit la procédure de Saikkonen pour l'estimation efficace du vecteur cointégrant  $\eta$ . Engle et Yoo (1991) proposèrent une autre approche. Elle implique une procédure d'estimation en trois étapes qui débute à partir des estimations en deux étapes de Engle-Granger et qui exploite une régression artificielle pour une étape de Gauss-Newton unique. D'autres auteurs, parmi lesquels Johansen (1988, 1991) et Phillips (1991a), ont proposé des méthodes d'estimation systémiques diverses. L'approche de Johansen sera exposée dans la section suivante.

Un grand nombre de travaux empiriques s'appuient sur des tests de cointégration et sur l'estimation de modèles avec des variables cointégrées. Des exemples de ces travaux sont Hall (1986), Baillie et Selover (1987), Campbell (1987), Campbell et Shiller (1987), Corbae et Ouliaris (1988), Granger et Lee (1989), Kunst et Neusser (1990), Johnson (1990), et King, Plosser, Stock, et Watson (1991). Une extension intéressante a été proposée au cas des séries temporelles saisonnières; voir Hylleberg, Engle, Granger, et Yoo (1990).

## 20.8 AUTORÉGRESSIONS VECTORIELLES ET COINTÉGRATION

L'une des approches les plus intéressantes à l'estimation systémique des modèles à variables cointégrées a été développée par Johansen (1988, 1991) et Johansen et Juselius (1990, 1992). Elle se base sur l'estimation d'une **autorégression vectorielle**, ou **VAR**, par la méthode du maximum de vraisemblance; voir la Section 19.5 pour davantage de détails sur les VAR. Dans cette section, nous discuterons brièvement de cette approche.

Considérons la VAR suivante avec un ensemble de variables en niveaux:

$$\mathbf{Y}_t = \mathbf{Y}_{t-1}\mathbf{\Pi}_1 + \cdots + \mathbf{Y}_{t-p}\mathbf{\Pi}_p + \mathbf{U}_t. \quad (20.33)$$

La notation est ici similaire à celle employée dans la Section 19.5:  $\mathbf{Y}_t$  et  $\mathbf{U}_t$  sont des vecteurs lignes de dimension  $1 \times m$ , et les matrices de dimension  $m \times m$   $\mathbf{\Pi}_1$  à  $\mathbf{\Pi}_p$  contiennent des coefficients. Par souci de simplicité, il n'y a pas de terme constant, bien que cette hypothèse soit rarement pertinente dans la réalité. On peut reparamétriser la VAR (20.33) comme suit:

$$\Delta \mathbf{Y}_t = \Delta \mathbf{Y}_{t-1}\mathbf{\Gamma}_1 + \cdots + \Delta \mathbf{Y}_{t-p+1}\mathbf{\Gamma}_{p-1} - \mathbf{Y}_{t-p}\mathbf{\Pi} + \mathbf{U}_t, \quad (20.34)$$

où  $\Gamma_1 = \Pi_1 - \mathbf{I}$ ,  $\Gamma_2 = \Pi_2 + \Gamma_1$ ,  $\Gamma_3 = \Pi_3 + \Gamma_2$ , et ainsi de suite. Ainsi la matrice  $\Pi$  est reliée aux matrices  $\Pi_i$  de (20.33) selon la formule

$$\Pi = \mathbf{I} - \Pi_1 - \cdots - \Pi_p.$$

En empilant les  $n$  observations de (20.34), nous obtenons le système complet

$$\Delta \mathbf{Y} = \Delta \mathbf{Y}_{-1} \Gamma_1 + \cdots + \Delta \mathbf{Y}_{-(p-1)} \Gamma_{p-1} - \mathbf{Y}_{-p} \Pi + \mathbf{U}, \quad (20.35)$$

où la notation ne nécessite aucun éclaircissement. Chaque terme de (20.35) est une matrice de dimension  $n \times m$ .

La matrice  $\Pi$ , que l'on appelle souvent **matrice d'impact**, détermine si oui ou non, et dans quelle mesure, le système (20.35) est cointégré. Si nous supposons comme d'habitude que les variables dont nous avons calculé les différences  $\Delta \mathbf{Y}$  sont stationnaires, alors chaque terme de (20.34) à l'exception de  $\mathbf{Y}_{t-p} \Pi$  est un élément d'un processus stationnaire. Cela implique la stationnarité de  $\mathbf{Y} \Pi$ . À l'évidence,  $\mathbf{Y} \Pi$  sera stationnaire si  $\Pi$  est une matrice composée d'éléments nuls. Ce doit être le cas lorsqu'aucune des séries n'est cointégrée avec une quelconque autre série. À l'autre extrême, si la matrice  $\Pi$  est de plein rang, seule la stationnarité de  $\mathbf{Y}$  implique celle de  $\mathbf{Y} \Pi$ , ce qui signifie que chacune des colonnes de  $\mathbf{Y}$  est stationnaire. Ces colonnes sont les différentes séries,  $\mathbf{y}_i$ ,  $i = 1, \dots, m$ , qui forment le système (20.33).

Entre ces deux positions radicales, si toutes les variables de  $\mathbf{Y}$  sont non stationnaires, (20.34) implique la cointégration, et que toute combinaison linéaire des colonnes de  $\mathbf{Y} \Pi$  doit être une série stationnaire. Supposons que  $\Pi$  soit de rang  $r$ , avec  $0 < r < m$ . Si c'est effectivement le cas, nous pouvons exprimer  $\Pi$  sous la forme

$$\Pi = -\boldsymbol{\eta} \boldsymbol{\alpha}^\top, \quad (20.36)$$

où  $\boldsymbol{\alpha}$  et  $\boldsymbol{\eta}$  sont des matrices de dimension  $m \times r$ , et où le signe négatif a été introduit par commodité. À partir de (20.36), nous voyons que  $\mathbf{Y}_{-p} \Pi = -\mathbf{Y}_{-p} \boldsymbol{\eta} \boldsymbol{\alpha}^\top$ . Les vecteurs cointégrants sont proportionnels aux colonnes de la matrice  $\boldsymbol{\eta}$ . Ainsi, pour chaque colonne de  $\boldsymbol{\eta}_i$ ,  $\mathbf{Y} \boldsymbol{\eta}_i$  est une variable aléatoire stationnaire. Lorsque  $r = 1$ , il n'existe qu'un unique vecteur cointégrant, et il est proportionnel à  $\boldsymbol{\eta}_1$ . Lorsque  $r = 2$ , il existe un espace bidimensionnel de vecteurs cointégrants, engendré par  $\boldsymbol{\eta}_1$  et  $\boldsymbol{\eta}_2$ , et ainsi de suite. Les deux cas extrêmes sont ceux pour lesquels  $r = 0$ , lorsqu'il n'existe aucun vecteur cointégrant, et  $r = m$ , lorsque toute combinaison linéaire des  $\mathbf{y}_i$  est stationnaire, parce que chaque  $\mathbf{y}_i$  est  $I(0)$ .

L'approche de Johansen (1988, 1991) consiste à estimer la VAR (20.34) soumise à la contrainte (20.36) pour des valeurs diverses de  $r$ , par maximum de vraisemblance. Cette estimation se base sur l'hypothèse que le vecteur d'aléas  $\mathbf{U}_t$  est normal multivarié pour tout  $t$  et indépendant des vecteurs d'aléas des autres observations. Cette hypothèse est moins contraignante qu'elle ne le paraît, puisqu'un nombre suffisamment grand de retards des différences de  $\mathbf{Y}$



dans (20.34) doit empêcher l'apparition d'une quelconque autocorrélation dans les résidus. Comme l'a montré Johansen, il est possible de maximiser la fonction de vraisemblance de manière analytique conditionnellement à n'importe quelle valeur de  $r$ , par une méthode similaire à celle employée dans la Section 18.5 pour obtenir des estimations LIML.

Le système (20.35) soumis à la contrainte (20.36) s'écrit

$$\Delta \mathbf{Y} = \Delta \mathbf{Y}_{-1} \mathbf{\Gamma}_1 + \cdots + \Delta \mathbf{Y}_{-(p-1)} \mathbf{\Gamma}_{p-1} + \mathbf{Y}_{-p} \boldsymbol{\eta} \boldsymbol{\alpha}^\top + \mathbf{U}. \quad (20.37)$$

Nous savons que les estimations ML des paramètres de ce système sont obtenues en minimisant le déterminant de la matrice des carrés et des produits croisés (souvenons-nous de la fonction de logvraisemblance concentrée (9.65)), c'est-à-dire

$$\left| \begin{aligned} & (\Delta \mathbf{Y} - \Delta \mathbf{Y}_{-1} \mathbf{\Gamma}_1 - \cdots - \Delta \mathbf{Y}_{-(p-1)} \mathbf{\Gamma}_{p-1} - \mathbf{Y}_{-p} \boldsymbol{\eta} \boldsymbol{\alpha}^\top)^\top \\ & (\Delta \mathbf{Y} - \Delta \mathbf{Y}_{-1} \mathbf{\Gamma}_1 - \cdots - \Delta \mathbf{Y}_{-(p-1)} \mathbf{\Gamma}_{p-1} - \mathbf{Y}_{-p} \boldsymbol{\eta} \boldsymbol{\alpha}^\top) \end{aligned} \right|.$$

On peut apercevoir à partir de cette expression que tous les éléments de  $\boldsymbol{\eta}$  et  $\boldsymbol{\alpha}$  ne peuvent pas être identifiés, puisque la factorisation (20.36) n'est pas unique pour une matrice  $\mathbf{\Pi}$  donnée. En fait, si  $\mathbf{B}$  est une matrice non singulière quelconque de dimension  $r \times r$ ,

$$\boldsymbol{\eta} \mathbf{B} \mathbf{B}^{-1} \boldsymbol{\alpha} = \boldsymbol{\eta} \boldsymbol{\alpha}.$$

Ainsi la matrice  $\boldsymbol{\eta}$  peut être élaborée en sélectionnant dans l'espace  $\mathcal{S}(\mathbf{\Pi})$  de dimension  $r$  n'importe quel ensemble de  $r$  vecteurs à  $m$  composantes linéairement indépendants. Une fois la matrice  $\boldsymbol{\eta}$  choisie,  $\boldsymbol{\alpha}$  est, de fait, uniquement déterminé. Cette propriété permet de contourner le problème de la dépendance non linéaire des fonctions de régression dans (20.37) vis-à-vis des paramètres.

On peut concentrer le déterminant par rapport aux paramètres des matrices  $\mathbf{\Gamma}_1$  à  $\mathbf{\Gamma}_{p-1}$  en les remplaçant par leurs estimations par moindres carrés. Ainsi, si nous notons  $\mathbf{M}_\Delta$  la projection orthogonale sur l'espace  $\mathcal{S}^\perp(\Delta \mathbf{Y}_{-1} \cdots \Delta \mathbf{Y}_{-(p-1)})$ , le déterminant qu'il s'agit de minimiser peut s'exprimer comme une fonction de  $\boldsymbol{\eta}$  et  $\boldsymbol{\alpha}$  uniquement, comme suit:

$$\left| (\Delta \mathbf{Y} - \mathbf{Y}_{-p} \boldsymbol{\eta} \boldsymbol{\alpha}^\top)^\top \mathbf{M}_\Delta (\Delta \mathbf{Y} - \mathbf{Y}_{-p} \boldsymbol{\eta} \boldsymbol{\alpha}^\top) \right|. \quad (20.38)$$

Si  $\mathbf{M}_\Delta \mathbf{Y}_{-p}$  désigne  $\mathbf{Y}_{-p}^*$ , et si  $\mathbf{M}_\Delta \mathbf{Y}$  désigne  $\Delta \mathbf{Y}^*$ , (20.38) peut s'écrire

$$\left| (\Delta \mathbf{Y}^* - \mathbf{Y}_{-p}^* \boldsymbol{\eta} \boldsymbol{\alpha}^\top)^\top (\Delta \mathbf{Y}^* - \mathbf{Y}_{-p}^* \boldsymbol{\eta} \boldsymbol{\alpha}^\top) \right|. \quad (20.39)$$

Il est désormais aisé de concentrer cette expression par rapport à  $\alpha$ , car, à condition de fixer  $\eta$ , les résidus dans (20.39) sont linéaires en  $\alpha$ . Si  $V \equiv Y_{-p}^* \eta$ , nous obtenons le déterminant

$$|(\Delta Y^*)^\top M_V \Delta Y^*|. \quad (20.40)$$

Par une astuce comparable à celle développée dans la Section 18.5, nous pouvons traiter (20.40) comme un seul facteur dans la décomposition du déterminant d'une matrice plus importante. Considérons

$$\begin{vmatrix} (\Delta Y^*)^\top \Delta Y^* & (\Delta Y^*)^\top V \\ V^\top \Delta Y^* & V^\top V \end{vmatrix}.$$

En exploitant le résultat (A.26) de l'Annexe A, cette matrice peut être factorisée soit comme

$$|V^\top V| |(\Delta Y^*)^\top M_V \Delta Y^*|$$

soit comme

$$|(\Delta Y^*)^\top \Delta Y^*| |V^\top M^* V|,$$

où  $M^*$  est la matrice de projection orthogonale associée à  $\mathcal{S}^\perp(\Delta Y^*)$ . Puisque  $|(\Delta Y^*)^\top \Delta Y^*|$  ne dépend pas de  $\eta$ , nous voyons que minimiser (20.40) est équivalent à minimiser le rapport

$$\frac{|V^\top M^* V|}{|V^\top V|} = \frac{|\eta^\top (Y_{-p}^*)^\top M^* Y_{-p}^* \eta|}{|\eta^\top (Y_{-p}^*)^\top Y_{-p}^* \eta|} \quad (20.41)$$

par rapport à  $\eta$ . Le minimum de (20.40) est alors celui de (20.41) multiplié par  $|(\Delta Y^*)^\top \Delta Y^*|$ .

La problème du ratio de moindre variance qu'il fallait résoudre dans le contexte LIML (voir (18.49)) faisait intervenir un rapport de formes quadratiques plutôt qu'un rapport de déterminants tel qu'il apparaît dans (20.41). Malgré cela, nous pouvons résoudre le problème par la même technique que (18.49), à savoir en le transformant en un problème de valeurs et de vecteurs propres. Avant de s'engager dans des détails, remarquons que (20.41) n'est pas modifié si nous remplaçons  $\eta$  par  $\eta B$ , pour toute matrice  $B$  de dimension  $r \times r$  non singulière. C'est précisément ce que nous relevions plus tôt en parlant de non unicité de (20.36). Nous ne pouvons donc pas espérer obtenir un unique  $\eta$ , mais au contraire tout un sous-espace de dimension  $r$ .

En ce qui concerne la minimisation présente, il est commode de se servir d'une transformation de  $\eta$ . Soit  $S$  n'importe quelle matrice de dimension  $m \times m$  telle que  $S^\top S = (Y_{-p}^*)^\top Y_{-p}^*$ , et définissons la matrice  $\zeta$  de dimension  $m \times r$  par  $S\eta$ . Le rapport (20.41) devient

$$\frac{|\zeta^\top (S^{-1})^\top (Y_{-p}^*)^\top M^* Y_{-p}^* S^{-1} \zeta|}{|\zeta^\top \zeta|}. \quad (20.42)$$

Puisque tout ce qui nous importe est le sous-espace engendré par les  $r$  colonnes de  $\zeta$ , nous pouvons choisir sans perte de généralité la matrice  $\zeta$  de telle sorte que  $\zeta^\top \zeta = \mathbf{I}_r$ . Soit  $\mathbf{A}$  la matrice définie positive de dimension  $m \times m$  qui apparaît au numérateur de (20.42). Il reste à minimiser  $|\zeta^\top \mathbf{A} \zeta|$  par rapport à  $\zeta$  sous la contrainte  $\zeta^\top \zeta = \mathbf{I}$ .

Pour mener à bien cette opération, il est plus facile de travailler sur le problème en termes de valeurs et vecteurs propres associés à  $\mathbf{A}$ . La résolution de ce problème nous fournira une matrice orthogonale  $\mathbf{Z}$ , dont les colonnes sont les vecteurs propres orthonormés de  $\mathbf{A}$ , et une matrice diagonale  $\mathbf{\Lambda}$ , dont les éléments diagonaux sont les valeurs propres de  $\mathbf{A}$ , qui doivent bien entendu être comprises entre 0 et 1. Alors  $\mathbf{AZ} = \mathbf{Z}\mathbf{\Lambda}$ . Si les colonnes de  $\mathbf{Z}$  et  $\mathbf{\Lambda}$  sont classées par ordre croissant des valeurs propres  $\lambda_1, \dots, \lambda_m$ , les estimations ML  $\hat{\zeta}$  peuvent être assimilées aux  $r$  premières colonnes de  $\mathbf{Z}$ . Géométriquement, les colonnes de  $\hat{\zeta}$  engendrent l'espace engendré par les vecteurs propres de  $\mathbf{A}$  qui correspondent aux  $r$  valeurs propres les plus petites. L'orthogonalité de  $\mathbf{Z}$  signifie que  $\hat{\zeta}$  satisfait la contrainte, et le choix des valeurs propres *les plus faibles* sert à minimiser le déterminant  $|\zeta^\top \mathbf{A} \zeta|$ .

On peut retrouver l'estimation ML de l'espace des vecteurs cointégrants  $\mathcal{S}(\eta)$  à partir de  $\hat{\zeta}$  grâce à la formule  $\hat{\eta} = \mathbf{S}^{-1} \hat{\zeta}$ . La matrice  $\hat{\alpha}$  requise pour l'obtention des estimations ML des paramètres de la matrice  $\Pi$  peut s'obtenir par la régression multivariée par OLS de  $\Delta \mathbf{Y}^*$  sur  $\mathbf{Y}_{-p}^* \hat{\eta}$ . Il en découle que les estimations des matrices  $\mathbf{I}_i$ ,  $i = 1, \dots, p-1$ , peut aussi s'obtenir par OLS.

Bien souvent, nous ne sommes pas particulièrement intéressés par les paramètres de la VAR (20.35). Notre préoccupation concerne davantage le test de l'hypothèse de non cointégration contre l'hypothèse alternative de cointégration d'un ordre quelconque. Si nous devons rejeter l'hypothèse nulle que  $r = 0$ , nous souhaiterions tester l'hypothèse nulle  $r = 1$  contre l'hypothèse alternative  $r = 2$ , et ainsi de suite. Les valeurs propres  $\lambda_i$ ,  $i = 1, \dots, m$ , procurent un moyen très pratique d'y parvenir, en termes d'un test du rapport de vraisemblance. Il est clair que si nous sélectionnons une valeur quelconque de  $r$ , le déterminant minimisé  $|\zeta^\top \mathbf{A} \zeta|$  est simplement le produit des  $r$  valeurs propres les plus faibles,  $\lambda_1 \cdots \lambda_r$ . Le minimum de (20.40) correspond à ce produit, multiplié par  $|(\Delta \mathbf{Y}^*)^\top \Delta \mathbf{Y}^*|$ . Si  $r = 0$ , le minimum de (20.40) est simplement ce dernier déterminant. Les rapports de vraisemblance pour les différentes valeurs de  $r$  sont par conséquent des produits de quelques-unes des valeurs propres, élevés à la puissance  $n/2$ ; souvenons-nous de (9.65). Si nous calculons les logarithmes et multiplions par 2 afin d'obtenir une statistique LR, nous aboutissons à  $-n$  fois le produit des logarithmes des valeurs propres concernées.

De façon générale, la statistique LR du test de l'hypothèse nulle  $r = r_1$ ,  $0 \leq r_1 < m$ , contre l'hypothèse alternative  $r = r_2$ ,  $r_1 < r_2 \leq m$ , est

$$LR = -n \sum_{i=r_1+1}^{r_2} \log \lambda_i. \quad (20.43)$$

Cette expression est évidemment l'analogie de la statistique LR (18.50) dans le contexte LIML. Cependant, elle n'aura pas la distribution asymptotique usuelle du  $\chi^2$ . Au lieu de cela, sous les différentes hypothèses nulles que l'on peut tester, les statistiques LR (20.43) auront des distributions asymptotiques non standards qui dépendent du nombre de "degrés de liberté"  $r_2 - r_1$  et de la présence ou non d'une constante ou d'une tendance linéaire dans la VAR. Ces distributions sont tabulées par simulation, pour un nombre limité de cas, par Johansen et Juselius (1990). On peut également réaliser des inférences sur les éléments des vecteurs cointégrants (normalisés de manière adéquate) aux moyens de statistiques LR conditionnellement à une certaine valeur de  $r$ ; ces statistiques auront alors une distribution asymptotique du  $\chi^2$  sous l'hypothèse nulle testée. C'est une propriété commode de l'approche VAR.

## 20.9 CONCLUSION

Nous avons vu dans ce chapitre que la théorie asymptotique pour les variables  $I(1)$  est très différente de la théorie asymptotique classique et avec laquelle nous sommes familiers. Du fait d'une différence aussi importante, nous n'avons pas tenté de la traiter trop en profondeur. Nous nous sommes contentés d'exposer quelques résultats fondamentaux de manière intuitive, et de fournir les références adéquates. La majeure partie des éléments présentés est relativement récente, à cause de l'effervescence théorique qui caractérise ce champ de recherches depuis une dizaine d'années, et une partie de ceux-ci est encore controversée. Les lecteurs peuvent aisément vérifier tout cela en lisant Phillips (1991b, 1991c) et d'autres articles chez Pesaran (1991).

## TERMES ET CONCEPTS

autorégression vectorielle (VAR)	tests de Dickey-Fuller augmentés
cointégration	(ADF)
erreurs d'équilibre	tests de Engle-Granger (EG)
estimateur super-convergent	tests de Engle-Granger augmentés
étendue (d'un ensemble de données)	(AEG)
matrice d'impact	tests de racine unitaire
méthode de Engle-Granger en deux	tests de racine unitaire non
étapes	paramétriques
processus de somme partielle	tests en $\tau$ , $\tau'$ , et $\tau^*$
processus de somme partielle	tests en $z$ et $z^*$
standardisé	théorèmes de la limite centrale
processus de Wiener standardisé	fonctionnels
racine unitaire	variables cointégrées
tests de cointégration sur résidus	vecteur cointégrant
tests de Dickey-Fuller (DF)	

# Chapitre 21

## Les Expériences Monte Carlo

### 21.1 INTRODUCTION

La plupart des méthodes d'estimation et de test d'hypothèse discutées dans ce livre ont des propriétés statistiques connues seulement asymptotiquement. Ceci est vrai pour les modèles non linéaires de tous types, pour les modèles d'équations simultanées linéaires, et même pour le modèle de régression linéaire univarié dès que nous relâchons l'hypothèse forte de régresseurs fixes ou l'hypothèse encore plus forte d'aléas normalement et identiquement distribués. Ainsi, dans la pratique, la théorie exacte en échantillon fini est rarement valable pour interpréter des estimations ou des statistiques de test. Malheureusement, à moins que la taille de l'échantillon ne soit effectivement très grande, il est très difficile de savoir si la théorie asymptotique est suffisamment précise pour nous permettre d'interpréter nos résultats en toute confiance.

Il existe fondamentalement deux manières de gérer cette situation. La première est d'affiner les approximations asymptotiques telles celles dérivées dans ce livre en additionnant des termes d'ordre inférieur par rapport à la taille de l'échantillon,  $n$ , termes qui sont typiquement  $O(n^{-1/2})$  ou  $O(n^{-1})$ . On fait référence à ces approximations plus raffinées en tant qu'**approximations en échantillon fini** ou **développements asymptotiques**. C'est l'étude des propriétés des estimateurs des modèles d'équations simultanées et des modèles dynamiques linéaires univariés qui a permis de décrire le plus largement l'approche des développements asymptotiques. Cette approche peut, dans certains cas, fournir des éclaircissements utiles sur le comportement des estimateurs et des statistiques de test. Malheureusement, elle implique souvent des éléments mathématiques soit plus avancés soit plus pénibles que ne le souhaiteraient la plupart des économètres. Cette méthode ne s'applique parfois qu'aux modèles relativement simples, et tend à produire des résultats compliqués et très difficiles à interpréter, en partie parce qu'ils dépendent souvent de paramètres inconnus. De plus, ces résultats ne sont eux-mêmes que de simples approximations; même s'ils sont généralement meilleurs que les approximations asymptotiques, ils peuvent ne pas être suffisamment précis. De façon idéale, on voudrait pouvoir utiliser automatiquement les développements asymptotiques, comme composante des applications de logiciels d'économétrie, afin d'obtenir des intervalles de confiance et des tests

d'hypothèses plus précis que ceux, asymptotiques, discutés dans ce livre. Malheureusement, cette situation idéale est peu fréquente, bien qu'un article récent de Rothenberg (1988) nous ait peut-être redonné un peu d'optimisme. Deux synthèses utiles des méthodes basées sur des développements asymptotiques sont Phillips (1983) et Rothenberg (1984). Une synthèse quelque peu critique de la littérature est Taylor (1983).

La seconde approche, que nous exposons dans ce chapitre, consiste à examiner les propriétés en échantillon fini des estimateurs et des statistiques de test en utilisant **les expériences Monte Carlo**. Le terme "Monte Carlo" est employé dans de nombreuses disciplines et fait référence aux procédures où les quantités d'intérêt sont approximées en générant de nombreuses réalisations aléatoires d'un processus stochastiques quelconque et en calculant une moyenne quelconque de leurs valeurs.<sup>1</sup> Puisque cela est pratiquement impossible à faire sans un ordinateur puissant, la littérature sur les **méthodes Monte Carlo** est assez récente. L'approche des développements asymptotiques nécessite une quantité de travail hautement qualifié très importante. Par contraste, l'approche Monte Carlo, comme Summers (1965) l'a souligné, est relativement intensive en capital. Elle économise du travail qualifié en consommant un temps de calcul sur ordinateur important.

Dans les applications économétriques des méthodes Monte Carlo, les grandeurs d'intérêt sont généralement des aspects variés des distributions des estimateurs et des statistiques de test, tels la moyenne et l'erreur quadratique moyenne d'un estimateur, le niveau d'une statistique de test sous l'hypothèse nulle, ou la puissance d'une statistique de test sous une hypothèse alternative quelconque. Hendry (1984) développe une étude provoquante. Cependant, la plus grande part de la littérature portant sur les méthodes Monte Carlo ne concerne pas spécifiquement la statistique ou l'économétrie mais également les méthodes d'approximation des intégrales multiples ou des systèmes non linéaires de simulation. Néanmoins, des références classiques telles que Hammersley et Handscomb (1964), Rubinstein (1981), Kalos et Whitlock (1986), Ripley (1987), et Lewis et Orav (1989) contiennent beaucoup d'éléments utiles.

Bien que les méthodes Monte Carlo soient souvent considérées comme une alternative à l'approche des développements asymptotiques, les deux approches doivent être plus justement considérées comme complémentaires. Tout comme les expériences Monte Carlo peuvent être utilisées pour valider des approximations asymptotiques, elles peuvent également être utilisées pour valider des approximations basées sur des développements asymptotiques. De plus, il existe de nombreuses situations où des développements asymptotiques peuvent s'utiliser pour analyser des cas spécifiques simples, tout en portant son attention sur des problèmes qui nécessitent un examen pour des cas plus généraux à l'aide d'expériences Monte Carlo. Cependant, puisque

<sup>1</sup> Le terme a pour initiateurs Metropolis et Ulam (1949). S'il avait été créé un tout petit peu plus tard, nous aurions pu parler de "méthode Las Vegas" à la place de "méthode Monte Carlo."

les développements asymptotiques dépassent l'objectif de ce livre, nous ne détaillerons pas davantage les manières de les utiliser conjointement aux méthodes Monte Carlo.

Un article qui utilise typiquement les méthodes Monte Carlo en statistique ou en économétrie présente des résultats à partir de plusieurs (peut-être nombreuses) expériences Monte Carlo reliées. Chaque expérience implique plusieurs éléments que le chercheur doit spécifier. Tout d'abord, il doit y avoir un modèle économétrique, et un ensemble d'estimateurs ou de statistiques de test associé au modèle. L'objet des expériences est d'examiner les propriétés en échantillon fini de ces estimateurs ou statistiques de test. Ensuite, il doit y avoir un processus générateur de données (DGP), qui est habituellement, mais pas toujours, un cas particulier du modèle. Le DGP doit être spécifié complètement. Ceci signifie que s'il y a des variables exogènes, elles ou leurs distributions doivent être spécifiées, comme doivent l'être les distributions de n'importe quel aléa. Chaque expérience se compose d'un nombre quelconque de **répétitions**, que nous noterons  $N$ . Chaque répétition implique de générer un seul ensemble de données à partir du DGP, et de calculer des estimateurs ou statistiques de test d'intérêt. Typiquement, le nombre de répétitions est très grand ( $N = 1000, 2000, 5000$ , et  $10,000$  sont des choix fréquents), mais il peut parfois être plus petit, par exemple  $50$ , si l'estimation prend beaucoup de temps et des résultats précis ne sont pas nécessaires. Après que  $N$  répétitions ont été opérées, on dispose de  $N$  observations sur chacun des estimateurs ou statistiques de test d'intérêt, et cet échantillon généré peut être soumis à l'analyse statistique pour calculer les estimations des quantités d'intérêt. Les résultats de l'expérience Monte Carlo sont ainsi eux-mêmes des estimations, et sont par conséquent associés à une erreur expérimentale. Cependant, nous pouvons minimiser cette erreur de façon acceptable en concevant avec soin l'expérience, en utilisant un nombre suffisamment grand de répétitions, et peut-être en appliquant des **techniques de réduction de variance** (consulter les Sections 21.5 et 21.6 qui suivent).

Comme la discussion précédente l'implique, il est rare de ne réaliser qu'une seule expérience Monte Carlo. En effet, les chercheurs exécutent généralement un ensemble d'expériences reliées, dans lequel la taille d'échantillon  $n$  et d'autres aspects du DGP (tels que les valeurs paramétriques) sont variés, afin de voir comment de telles variations affectent les estimateurs ou statistiques de test d'intérêt. S'il n'y a que quelques expériences, les résultats sont habituellement présentés sous la forme d'un tableau. Cependant, s'il y a de nombreuses expériences, ce tableau peut comporter un très grand nombre d'éléments, que les lecteurs peuvent juger difficile à assimiler. Une manière de traiter un tel problème est d'estimer une **surface de réponse**, où les résultats de chaque expérience sont traités comme une seule observation, et un modèle de régression ajuste les quantités d'intérêt à la taille d'échantillon et aux autres aspects du DGP qui varient selon l'expérience. De façon idéale, les estimations de la surface de réponse résument les résultats des expériences et fournissent

une manière plus compacte et plus rapidement compréhensible de présenter les résultats qu'une suite de tableaux ne le ferait. L'approche de la surface de réponse sera discutée dans la Section 21.7.

Dans la suite de ce chapitre, nous discutons des caractéristiques importantes des expériences Monte Carlo en économétrie. La plupart des expériences Monte Carlo nécessitent un grand nombre de **variables pseudo-aléatoires**, c'est-à-dire de nombres qui semblent être des tirages d'une distribution de probabilité spécifiée quelconque. Dans les deux prochaines sections, nous discutons brièvement de la façon de générer ces nombres sur ordinateur. Dans la Section 21.4, nous aborderons d'autres aspects de conception d'un ensemble d'expériences Monte Carlo. Dans les Sections 21.5 et 21.6, nous discutons des techniques de réduction de variance, qui sont souvent utilisées pour accroître la précision des résultats pour un temps de calcul imparti. Dans la section suivante, nous parlons de l'utilisation des surfaces de réponse. Enfin, dans la Section 21.8, nous discutons brièvement de la méthode statistique connue sous le nom de **bootstrap**, qui est très étroitement reliée aux méthodes Monte Carlo.

## 21.2 GÉNÉRATION DES NOMBRES PSEUDO-ALÉATOIRES

Chaque expérience Monte Carlo nécessite un grand nombre de variables "aléatoires", issues d'une ou plusieurs distributions préspecifiées. Par exemple, considérons une petite expérience traitant d'un modèle de régression comportant des régresseurs fixes. Supposons qu'il y ait 50 observations pour 1000 répétitions. Pour une telle expérience, un total de 50,000 variables "aléatoires" serait nécessaire simplement pour générer les aléas. S'il y avait trois régresseurs stochastiques, un complément de 150,000 variables "aléatoires" serait nécessaire pour générer les régresseurs. Comme nous le verrons dans la prochaine section, si nous pouvons trouver une manière d'obtenir des nombres "aléatoires" uniformément distribués sur l'intervalle 0-1, noté  $U(0, 1)$ , il est alors habituellement très facile d'obtenir des variables "aléatoires" distribuées selon n'importe quelle distribution que nous spécifions. Le problème fondamental consiste à obtenir les nombres "aléatoires" initiaux. Bien qu'il soit possible d'acquérir de façon *authentique* des nombres aléatoires au moyen d'observations physiques telles que la décomposition des isotopes radioactifs, il serait extrêmement inconfortable de connecter son ordinateur à un générateur de nombres physiques aléatoires, ou de lui faire lire un tableau immense de nombres aléatoires collectés au préalable, à chaque fois que nous avons à exécuter une expérience Monte Carlo! Ainsi, il est évident que si les expériences Monte Carlo doivent être pratiques, il faut que l'ordinateur puisse générer de manière autonome, rapidement et à moindre coût des nombres "aléatoires".

Dans le paragraphe précédent, les guillemets autour du mot "aléatoire" insistaient sur le fait que ce dont nous avons besoin, pour lancer une expérience



Monte Carlo, c'est une manière d'obtenir des nombres qui possèdent les mêmes propriétés statistiques que des nombres aléatoires, plutôt que des nombres véritablement aléatoires. En effet, aucun ordinateur n'est capable de générer des nombres aléatoires authentiques, du moins pas s'il travaille correctement. Mais les ordinateurs sont capables de générer des suites de **nombres pseudo-aléatoires**, qui sont en fait purement déterministes. Les programmes qui procèdent ainsi sont appelés **générateurs de nombres pseudo-aléatoires** ou, plus communément mais de façon moins précise, simplement **générateurs de nombres aléatoires**. Les nombres pseudo-aléatoires générés par un générateur de nombres aléatoires performant sont, pour nos objectifs des expériences Monte Carlo, indiscernables des suites de nombres aléatoires authentiques, c'est-à-dire de véritables suites de tirages indépendants issus de la distribution  $U(0, 1)$ .

Il existe de nombreuses manières de générer des nombres pseudo-aléatoires. Les plus communes sont des variantes du **générateur congruentiel**,

$$\eta_t = \frac{z_t}{m}, \quad z_t = (\lambda z_{t-1} + \alpha)(\text{mod } m), \quad (21.01)$$

où  $\eta_t$  est le  $i^{\text{ième}}$  nombre aléatoire généré, et  $z_t$  est un entier positif. Le générateur (21.01) dépend de trois paramètres:  $\lambda$  est appelé **multiplicateur**,  $\alpha$  l'**incrément**, et  $m$  le **module**. La notation  $(\text{mod } m)$  signifie que nous divisons ce qui la précède par  $m$  et retenons le reste. Ainsi,  $z_t$  doit être inférieur à  $m$ , et  $\eta_t$  doit toujours être compris entre 0 et 1. Nous pouvons montrer qu'un générateur congruentiel doit toujours se répéter en fin de compte, dans au plus  $m$  étapes, de sorte que nous sélectionnerons un  $m$  aussi grand que possible. Par conséquent,  $m$  prend souvent la valeur du plus grand entier qui peut être représenté de façon exacte par un ordinateur particulier; il s'agit fréquemment de  $2^{31} - 1$ . Avec ce choix de  $m$ , nous pourrions, en principe, générer quelque chose comme plus de deux milliards de nombres aléatoires avant que la suite ne se répète. Cependant, si  $m$ ,  $\lambda$ , et  $\alpha$  sont mal choisis, la suite peut se répéter plus rapidement et peut présenter d'autres symptômes de non stochasticité.

Le choix de l'incrément  $\alpha$  n'est pas si important; une variante largement utilisée de (21.01) est la classe des **générateurs congruentiels multiplicatifs**, où  $\alpha$  est nul. Cependant, le choix du multiplicateur  $\lambda$  est extrêmement important. Certains choix sont connus pour mener à des générateurs dont le comportement est relativement bon, tandis que d'autres sont connus pour conduire à de très mauvais générateurs. Pour plus de détails, consulter Kennedy et Gentle (1980), Knuth (1981), Rubinstein (1981), Press, Flannery, Teukolsky, et Vetterling (1986), Ripley (1987), L'Ecuyer (1988), et Lewis et Orav (1989).

La plupart du temps, les économètres qui effectuent une expérience Monte Carlo n'auront pas besoin d'écrire leurs propres générateurs de nombres aléatoires. S'ils utilisent un générateur efficace et de grande qualité, la seule chose dont ils doivent se soucier est de savoir comment se procurer la **valeur**

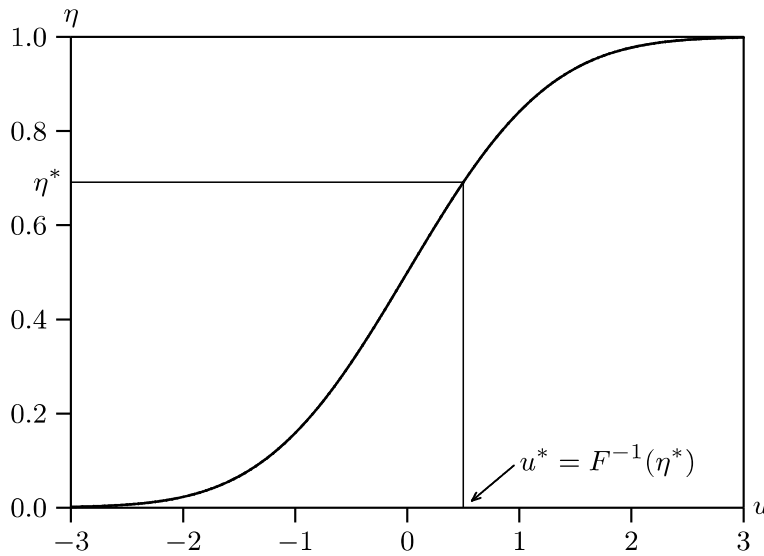
**d'origine**, (**seed** (graine) en anglais) qui est la valeur initiale  $z_0$  nécessaire pour générer  $z_1$  et qui, pour un générateur donné, détermine de façon unique la suite entière des nombres aléatoires. La valeur d'origine peut être spécifiée plus ou moins arbitrairement comme une valeur entière élevée inférieure à  $m$ , ou choisie "de façon aléatoire" à l'aide de l'horloge du système. Quel que soit le cas, elle devrait être enregistrée pour pouvoir répéter une expérience si nécessaire. La valeur d'origine n'est fournie que lorsque le générateur est lancé à partir d'un programme particulier. Après la première boucle,  $z_0$  est remplacée par  $z_1$ , ensuite par  $z_2$ , et ainsi de suite. Donc, à chaque fois, le programme stocke la valeur  $z_{t-1}$  pour calculer  $z_t$ .

Malheureusement, dans la réalité et l'utilisation courante, les générateurs de nombres aléatoires de mauvaise qualité sont nombreux, et il est sûrement imprudent de se fier à un générateur qui n'a pas subi des tests variés. De tels tests sont discutés dans la plupart des livres traitant des méthodes Monte Carlo mentionnés auparavant; consulter aussi Fishman et Moore (1982). Les tests que l'on souhaiterait exécuter dépendent de l'usage des nombres aléatoires. Si le modèle étudié est un modèle de série temporelle, par exemple, on voudrait être sûr qu'ils sont non soumis à une autocorrélation. Notons que les mauvais générateurs de nombres aléatoires peuvent souvent être améliorés en "mélangeant" des nombres qu'ils produisent ou en combinant plusieurs programmes d'une manière quelconque. Par exemple, nous pourrions utiliser deux programmes différents pour générer deux nombres aléatoires différents, puis utiliser un troisième programme pour déterminer de façon aléatoire lequel des deux choisir.

### 21.3 GÉNÉRER DES VARIABLES PSEUDO-ALÉATOIRES

Une fois que l'on dispose d'un programme pouvant générer de longues suites de nombres pseudo-aléatoires  $\eta_t$ , chacun étant apparemment distribué de façon indépendante suivant une  $U(0, 1)$ , les manières de générer des variables pseudo-aléatoires qui apparaissent être des tirages de n'importe quelle distribution désirée sont nombreuses. Nous examinerons deux techniques générales, la **méthode de transformation** et la **méthode de rejet**, ainsi que des méthodes spéciales variées qui s'appliquent à certains cas intéressants.

La méthode de transformation est basée sur le fait que l'espace d'arrivée d'une fonction de répartition (c.d.f.) est l'intervalle 0-1. Ainsi, si  $u$  est distribuée selon la c.d.f. strictement croissante  $F(u)$ ,  $\eta = F(u)$  doit être distribuée selon  $U(0, 1)$ . Pour tout  $\eta$ , nous pouvons inverser la c.d.f. et obtenir  $u = F^{-1}(\eta)$ . Pour obtenir une suite de  $u_t$  distribuées selon  $F(u)$ , nous générons simplement une suite de  $\eta_t$  distribuées selon  $U(0, 1)$  et soumettons chaque terme à la transformation  $F^{-1}(\eta_t)$ . C'est ce que montre la Figure 21.1. Comme nous le voyons d'après la figure, n'importe quelle valeur de  $\eta$  sur l'axe vertical, telle que  $\eta^*$ , est appliquée de façon unique par  $F^{-1}(\eta^*)$  à une valeur correspondante  $u^*$  sur l'axe horizontal.



**Figure 21.1** La méthode de transformation

La méthode de transformation fonctionne bien lorsque  $F^{-1}(\cdot)$  n'est pas difficile à calculer. C'est le cas avec la distribution exponentielle, dont la fonction de densité de probabilité (p.d.f.) est

$$f(u) = \theta e^{-\theta u}$$

(consulter la Section 8.1), et la c.d.f. correspondante est

$$F(u) = 1 - e^{-\theta u}.$$

Si nous posons  $\eta$  égale à  $F(u)$  et résolvons, nous trouvons que

$$u = F^{-1}(\eta) = -\frac{1}{\theta} \log(1 - \eta).$$

Ainsi, dans ce cas, la méthode de transformation peut facilement être utilisée pour générer des variables pseudo-aléatoires distribuées selon la distribution exponentielle.

La méthode de transformation peut être employée pour générer des variables pseudo-aléatoires normales, mais elle nécessite une certaine masse de calculs parce qu'il n'existe aucune expression formelle proche de la c.d.f. de la normale centrée réduite  $\Phi(\cdot)$  ou de son inverse  $\Phi^{-1}(\cdot)$ . On utilise un algorithme pour calculer numériquement cette dernière. Une technique alternative largement utilisée est la **méthode de Box-Muller** de Box et Muller (1958). Elle utilise le fait que si  $\eta_1$  et  $\eta_2$  sont des variables aléatoires indépendantes issues de  $U(0, 1)$ , alors les termes

$$u_1 = (-2 \log(\eta_1))^{1/2} \cos(2\pi\eta_2) \quad \text{et} \quad u_2 = (-2 \log(\eta_1))^{1/2} \sin(2\pi\eta_2)$$

sont des variables aléatoires indépendantes issues de  $N(0, 1)$ . Consulter Rubinstein (1981) ou Press, Flannery, Teukolsky, et Vetterling (1986) pour une démonstration. Le dernier livre discute également d'une version modifiée de la méthode de Box-Muller qui devrait être plus rapide à calculer. Le problème majeur avec la technique de Box-Muller est qu'elle repose fortement sur l'indépendance de  $\eta_1$  et  $\eta_2$ . Si le générateur de nombres aléatoires qui les produit n'est pas bon, ces variables peuvent manifester une certaine dépendance, et les variables résultantes  $u_1$  et  $u_2$  peuvent ne pas être normales ou indépendantes.

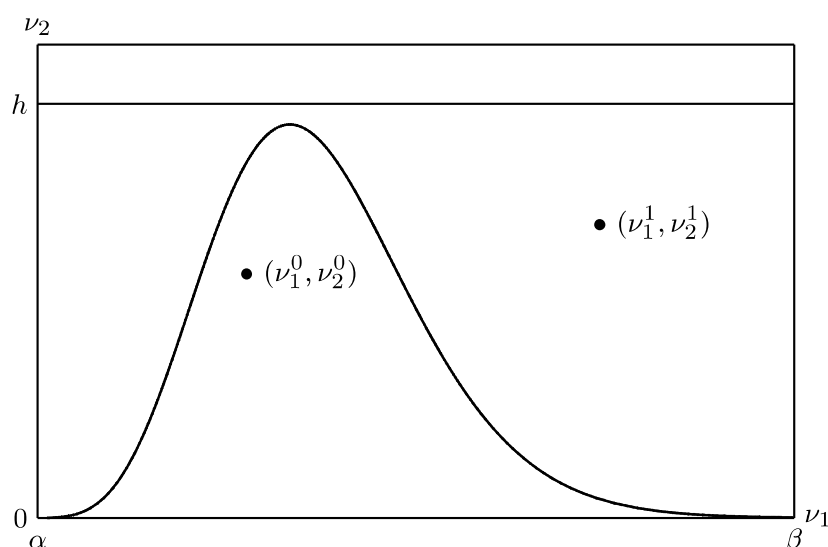
Si l'on est capable d'obtenir des variables pseudo-aléatoires à partir de  $N(0, 1)$ , il est immédiat d'obtenir des variables pseudo-aléatoires à partir de  $N(\mu, \sigma^2)$  ou à partir de la distribution normale multivariée avec n'importe quel vecteur d'espérances  $\mu$  et matrice de covariance  $\Omega$ . Si  $u$  désigne un vecteur de dimension  $l$  dont chaque élément est une variable pseudo-aléatoire issue de  $N(0, 1)$ , et si  $\psi$  est une matrice de dimension  $l \times l$  (habituellement triangulaire) telle que  $\psi^\top \psi = \Omega$ , il est facile de voir que le vecteur  $v$  de dimension  $l$  défini par

$$v \equiv \mu + \psi^\top u$$

suivra la distribution  $N(\mu, \Omega)$ . Des variables issues des distributions de Cauchy, du chi-carré, du  $t$  de Student, de Fisher, sont également immédiatement disponibles en exploitant simplement les relations entre ces distributions et la distribution normale centrée réduite, et entre chacune d'elles (consulter l'Annexe B). Par exemple, pour générer des variables aléatoires issues de  $\chi^2(5)$ , nous pourrions générer 5 variables aléatoires à partir de  $N(0, 1)$ , les mettre au carré, et sommer leurs carrés. Cette méthode fonctionne bien tant que le nombre de degrés de liberté est faible mais elle ne serait pas recommandée pour générer des variables aléatoires à partir de, disons,  $F(65, 1743)$ .

L'autre méthode fréquemment utilisée et largement applicable pour générer des variables aléatoires est la méthode de rejet. Elle peut s'utiliser chaque fois que la p.d.f.  $f(u)$  est connue. Dans sa version la plus simple, la méthode de rejet nécessite que l'espace de départ de  $f(u)$  soit un intervalle fini de la droite réelle, disons l'intervalle  $[\alpha, \beta]$ . On commence par obtenir deux variables aléatoires à partir de  $U(0, 1)$ , disons  $\eta_1$  et  $\eta_2$ . La première est transformée en  $\nu_1$ , une variable aléatoire provenant de  $U(\alpha, \beta)$ , tandis que la seconde est transformée en  $\nu_2$ , une variable aléatoire provenant de  $U(0, h)$ , où  $h$  est un nombre au moins aussi grand que le maximum de  $f(u)$ . Une fois obtenues  $\nu_1$  et  $\nu_2$ ,  $\nu_2$  est comparée à  $f(\nu_1)$ . Si  $\nu_2$  excède  $f(\nu_1)$ , la variable aléatoire proposée  $\nu_1$  est rejetée et une autre paire  $(\nu_1, \nu_2)$  est tirée de la distribution. Cependant, si  $\nu_2$  est inférieure ou égale à  $f(\nu_1)$ ,  $\nu_1$  est acceptée et  $u$  lui est égale. Cette méthode est illustrée dans la Figure 21.2. Ici le point  $(\nu_1^0, \nu_2^0)$  fournit une valeur  $u$ , tandis que le point  $(\nu_1^1, \nu_2^1)$  est rejeté.

Il est facile de voir pourquoi la méthode de rejet fonctionne correctement. Bien que nous extrayions  $\nu_1$  initialement à partir de  $U(\alpha, \beta)$ , nous l'acceptons



**Figure 21.2** La méthode de rejet

seulement si  $\nu_2 < f(\nu_1)$ , et la probabilité que ceci survienne est proportionnelle à  $f(\nu_1)$ . Cette version de la méthode de rejet est naturellement quelque peu inefficace, puisque nous devons générer, en moyenne,  $2h(\beta - \alpha)$  variables aléatoires pour chaque  $u$  que nous obtenons réellement. Si la densité  $f(u)$  a un sommet élevé,  $h$  sera grande. Si la densité a de longues queues,  $\beta - \alpha$  sera grand. Quel que soit le cas,  $2h(\beta - \alpha)$  sera grand, et la méthode peut se révéler relativement inefficace. Dans une version plus générale de la méthode de rejet, la constante  $h$  est remplacée par une fonction  $h(\nu_1)$ ,  $\nu_1$  étant alors issu d'une densité proportionnelle à  $h(\nu_1)$ . Alors on peut assimiler  $\nu_2$  à  $U(0, h(\nu_1))$ . Pourvu que  $h(\nu_1) > f(\nu_1)$  partout sur  $[\alpha, \beta]$ , qui n'est plus forcément fini, cette méthode est valable; pourvu qu'il soit facile de générer des variables aléatoires  $\nu_1$  avec une probabilité proportionnelle à  $h(\nu_1)$ , et que l'aire sous  $h(\cdot)$  ne soit pas beaucoup plus grande que l'aire sous  $f(\cdot)$ , elle fonctionnera efficacement. Notons que  $h(\cdot)$  n'est pas à proprement parler une densité, puisque  $h(\nu_1)$  doit être supérieure à  $f(\nu_1)$  pour tout  $\nu_1$  et par conséquent doit avoir une intégrale supérieure à l'unité; cependant, il peut être commode de sélectionner une fonction  $h(\cdot)$  proportionnelle à une densité bien connue quelconque.

## 21.4 CONCEPTION DES EXPÉRIENCES MONTE CARLO

L'étape la plus délicate pour réaliser un ensemble d'expériences Monte Carlo consiste habituellement à les concevoir. Les limites des possibilités de calcul, le temps disponible de expérimentateur, et la quantité d'espace que l'on peut raisonnablement consacrer à la présentation des résultats expliquent qu'il est habituellement pratique d'exécuter seulement un petit nombre d'expériences.

Celles-ci doivent être conçues pour apporter autant d'information que possible sur les problèmes qui nous intéressent.

La première chose à reconnaître est que les résultats issus des expériences Monte Carlo sont nécessairement aléatoires. Au minimum, cela signifie que les résultats doivent être exposés de telle manière que le lecteur apprécie l'étendue du hasard expérimental. De plus, il est essentiel d'exécuter suffisamment de répétitions pour que les résultats soient suffisamment précis pour le propos étudié. Le nombre de répétitions nécessaire peut parfois être réduit de façon substantielle en utilisant des techniques de réduction de variance dont nous discuterons dans les deux prochaines sections. Cependant de telles techniques ne sont pas toujours immédiatement disponibles. Dans cette section, nous considérons d'autres aspects variés de la conception des expériences Monte Carlo.

Nous considérons tout d'abord le problème qui consiste à déterminer combien de répétitions exécuter. Par exemple, supposons que le chercheur soit intéressé par le calcul du niveau d'une certaine statistique de test (c'est-à-dire la probabilité de rejet de l'hypothèse nulle quand elle est vraie), disons, au niveau nominal .05. Notons  $p$  cette quantité inconnue. Chaque répétition générera une statistique de test qui excède ou pas la valeur critique nominale. Celles-ci peuvent être assimilées à des tirages indépendants de la loi de Bernoulli. Supposons que  $N$  répétitions soient exécutées et  $R$  rejets obtenus. Alors l'estimateur évident de  $p$ , qui est aussi l'estimateur ML, est  $R/N$ . La variance de l'estimateur est  $N^{-1}p(1-p)$ , et peut être estimée par  $R(N-R)/N^3$ .

Supposons maintenant que l'on veuille que la longueur d'un intervalle de confiance à 95% sur l'estimation de  $p$  soit approximativement .01. En utilisant l'approximation normale de la binomiale, qui est ici sûrement valable puisque  $N$  sera grand, nous voyons que l'intervalle de confiance doit s'étendre sur  $2 \times 1.96 = 3.92$  écarts types. Par conséquent, nous avons besoin que

$$3.92 \left( \frac{p(1-p)}{N} \right)^{1/2} = .01. \quad (21.02)$$

En supposant que  $p$  soit .05, le niveau nominal du test étudié, nous pouvons trouver la valeur de  $N$  en résolvant (21.02). Le résultat est  $N \cong 7299$ . Pour prendre toutes les sécurités (puisque  $p$  peut bien excéder .05, impliquant une forte variance pour  $R/N$ ), le chercheur choisirait probablement  $N = 8000$ . Il s'agit d'un nombre plutôt grand de répétitions et il peut être très coûteux à calculer. Si l'on désire laisser la longueur de l'intervalle de confiance à 95% de  $p$  à .02, on pourrait sélectionner un échantillon réduit au quart, ou approximativement à 2000 répétitions.

Si l'objet d'une expérience est de comparer deux ou plusieurs estimateurs, ou deux ou plusieurs statistiques de test, un nombre plus petit de répétitions est nécessaire pour obtenir un niveau donné de précision par rapport à ce qui

serait nécessaire pour estimer les propriétés de ces estimateurs ou statistiques, à niveau de précision identique. Supposons, par exemple, que l'on veuille comparer deux estimateurs, disons  $\hat{\theta}$  et  $\tilde{\theta}$ , d'un paramètre dont la véritable valeur est  $\theta_0$ . A chaque répétition, disons la  $j^{\text{ième}}$ , les réalisations de chacun des deux estimateurs, disons  $\hat{\theta}_j$  et  $\tilde{\theta}_j$ , sont obtenues. Les écarts types des deux estimateurs sont

$$B(\hat{\theta}) \equiv E(\hat{\theta} - \theta_0) \quad \text{et} \quad B(\tilde{\theta}) \equiv E(\tilde{\theta} - \theta_0),$$

et peuvent être estimés par

$$\hat{B}(\hat{\theta}) = \frac{1}{N} \sum_{j=1}^N (\hat{\theta}_j - \theta_0) \quad \text{et} \quad \tilde{B}(\tilde{\theta}) = \frac{1}{N} \sum_{j=1}^N (\tilde{\theta}_j - \theta_0).$$

La différence entre  $B(\hat{\theta})$  et  $B(\tilde{\theta})$  est

$$E(\hat{\theta} - \theta_0) - E(\tilde{\theta} - \theta_0) = E(\hat{\theta} - \tilde{\theta}), \quad (21.03)$$

que l'on peut estimer par

$$\frac{1}{N} \sum_{j=1}^N (\hat{\theta}_j - \tilde{\theta}_j). \quad (21.04)$$

Il est possible et en effet vraisemblable que la variance de (21.04) sera substantiellement inférieure à la variance de  $\hat{B}(\hat{\theta})$  ou de  $\tilde{B}(\tilde{\theta})$ , parce que  $\hat{\theta}_j$  et  $\tilde{\theta}_j$  dépendent du même vecteur pseudo-aléatoire  $\mathbf{u}^j$ . La variance de (21.04) est

$$\frac{1}{N} V(\hat{\theta}) + \frac{1}{N} V(\tilde{\theta}) - \frac{2}{N} \text{Cov}(\hat{\theta}, \tilde{\theta}),$$

qui sera inférieure à la variance de  $\hat{B}(\hat{\theta})$  ou de  $\tilde{B}(\tilde{\theta})$  lorsque  $\text{Cov}(\hat{\theta}, \tilde{\theta})$  est positive et suffisamment grande. Ceci sera très souvent le cas, puisqu'il est très probable que  $\hat{\theta}_j$  et  $\tilde{\theta}_j$  soient fortement positivement corrélés. Ainsi, beaucoup moins de répétitions sont nécessaires pour estimer (21.03) que pour estimer  $B(\hat{\theta})$  et  $B(\tilde{\theta})$  à niveau de précision identique. Naturellement, ceci surviendra seulement si  $\hat{\theta}_j$  et  $\tilde{\theta}_j$  sont obtenues avec le même ensemble de variables pseudo-aléatoires, mais c'est exactement comme cela que l'expérience Monte Carlo serait conçue. Nous rencontrerons une idée similaire à celle-ci lorsque nous discuterons de la méthode des variables antithétiques dans la prochaine section.

La seconde chose importante à garder à l'esprit quand on conçoit des expériences Monte Carlo est que les résultats seront souvent très sensibles à certains aspects de la conception expérimentale mais pratiquement ou totalement insensibles à d'autres aspects. Evidemment, on voudra faire varier les premiers à travers les expériences tout en fixant les derniers d'une manière plus ou moins arbitraire. Par exemple, de nombreuses statistiques de test

reliées aux modèles de régression sont invariantes à la variance des aléas. Considérons le  $t$  de Student pour  $\alpha = 0$  dans la régression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{z} + \mathbf{u}. \quad (21.05)$$

En utilisant le Théorème FWL et en supposant que les données sont générées par un cas particulier de (21.05) pour lequel  $\alpha = 0$ , nous voyons que

$$t(\hat{\alpha}) = \frac{\mathbf{z}^\top \mathbf{M}_X \mathbf{u}}{(\mathbf{u}^\top \mathbf{M}_{X,z} \mathbf{u} / (n - k))^{1/2} (\mathbf{z}^\top \mathbf{M}_X \mathbf{z})^{1/2}}, \quad (21.06)$$

où il y a  $n$  observations et un total de  $k$  régresseurs et, comme d'habitude,  $\mathbf{M}_X$  et  $\mathbf{M}_{X,z}$  désignent les matrices qui projettent orthogonalement sur les sous-espaces  $\mathcal{S}^\perp(\mathbf{X})$  et  $\mathcal{S}^\perp(\mathbf{X}, \mathbf{z})$ , respectivement. La distribution en échantillon fini de cette statistique de test quand les  $u_t$  ne sont pas normaux est généralement inconnue et pourrait bien être le sujet d'une expérience Monte Carlo. Cependant, il est clair à partir de l'inspection de (21.06) que cette distribution ne dépend en aucune manière de la variance des aléas qui composent le vecteur d'aléas  $\mathbf{u}$  dans le DGP, puisque si nous multiplions  $\mathbf{u}$  par une constante positive quelconque,  $t(\hat{\alpha})$  est inchangé. Ainsi, dans ce cas, nous pourrions tout aussi bien fixer la variance des aléas à une certaine valeur arbitraire, puisqu'il n'y aurait rien du tout à apprendre en la faisant varier. Breusch (1980) discute d'un certain nombre d'autres résultats d'invariance pour des modèles de régression linéaire; en prenant en compte de tels résultats, on peut simplifier dans de nombreux cas la conception des expériences Monte Carlo.

Par ailleurs, quand il y a une raison de s'attendre à ce que les résultats soient sensibles à certains aspects du DGP, il est important de mener des expériences dans lesquelles ces aspects varient dans toute la gamme des aspects intéressants. Ces aspects du DGP qu'il faut faire varier doivent nécessairement être différents selon les cas. La taille d'échantillon  $n$  sera typiquement l'un d'eux, parce qu'il est presque toujours intéressant de voir avec quelle rapidité les propriétés en échantillon fini des quantités examinées approchent leurs limites asymptotiques (connues). Une exception à cela est le cas où le but de l'expérience Monte Carlo est de détailler les propriétés d'un ensemble particulier d'estimateurs ou de statistiques de test pour un ensemble de données particulier, de telle sorte que l'expérience est utilisée comme complément d'une partie d'un travail empirique (consulter la Section 21.8). Par contraste avec cette situation, jusqu'à présent la plupart des expériences Monte Carlo ont été conçues pour détailler les propriétés générales de certaines procédures statistiques, et il est difficile d'imposer n'importe quelle sorte de généralité quand tous les résultats sont relatifs à une seule taille d'échantillon.

La grande majorité des modèles qu'estiment les économètres consiste en des modèles de régression ou des modèles proches des modèles de régression. Ainsi, sauf dans quelques cas particuliers tels que les modèles chronologiques purs, des variables conditionnantes ( $\mathbf{X}_t$ ) sont habituellement présentes. La



manière dont celles-ci devraient être traitées dans les expériences Monte Carlo n'est pas vraiment claire. Une approche consiste à générer les  $\mathbf{X}_t$  d'une certaine manière. Lorsque l'expérience traite des données en coupe transversale, il est plus pratique de les générer à partir des distributions indépendantes des lois uniforme, normale ou lognormale, alors que lorsque l'expérience traite des données chronologiques, il est pratique de les générer à partir de processus variés simples de série temporelles tels que AR(1), MA(1), et ARMA(1,1), à aléas normaux. On peut soit générer un nouvel ensemble de  $\mathbf{X}_t$  pour chaque répétition soit générer un seul ensemble de  $\mathbf{X}_t$  utilisé dans toutes les répétitions. La dernière méthode est moins coûteuse et se justifie si les  $\mathbf{X}_t$  sont supposés fixes dans les échantillons répétés, mais elle peut conduire à des résultats qui dépendent des caractéristiques particulières de l'ensemble particulier des  $\mathbf{X}_t$  généré.

Une autre possibilité consiste à utiliser de véritables données économiques pour les  $\mathbf{X}_t$ . Si ces données sont choisies avec soin, cette approche peut garantir que les  $\mathbf{X}_t$  sont en fait typiquement celles qui apparaissent dans les modèles économétriques. Cependant, cela pose le problème de la variation de la taille d'échantillon. Si l'on utilise soit des données authentiques soit un seul ensemble de données générées, la matrice  $n^{-1}\mathbf{X}^\top\mathbf{X}$  variera avec la taille de l'échantillon  $n$ . Ceci peut rendre la distinction des effets des variations de  $n$  des effets des variations de  $n^{-1}\mathbf{X}^\top\mathbf{X}$  difficile. Une solution à ce problème est de sélectionner, ou de générer, un seul ensemble de  $\mathbf{X}_t$  pour un échantillon de taille  $m$  et de répéter ensuite ceux-ci autant de fois que nécessaire pour créer les  $\mathbf{X}_t$  pour les échantillons de tailles plus grandes. Ceci nécessite que  $n = cm$ , où  $c$  est un entier. Des choix évidents pour  $m$  sont 50 et 100;  $n$  pourrait ensuite être un entier quelconque multiple de 50 ou de 100. Naturellement, le problème avec cette approche est que comme beaucoup de répétitions sont exécutées, tous les résultats dépendront du choix de l'ensemble initial des  $\mathbf{X}_t$ .

Dans de nombreux cas, la manière de choisir les  $\mathbf{X}_t$  ne sera pas d'une grande importance. Cependant, il existe des cas pour lesquels elle peut avoir un impact substantiel sur les résultats. Par exemple, MacKinnon et White (1985) ont utilisé les expériences Monte Carlo pour examiner la performance en échantillon fini de différents estimateurs des matrices de covariance robustes à l'hétéroscédasticité (HCCME; consulter la Section 16.3). Ils ont utilisé 50 observations sur de véritables données économiques pour les  $\mathbf{X}_t$ , répétant ces 50 observations autant que nécessaire pour chaque taille d'échantillon. Comme Chesher et Jewitt (1987) l'ont montré plus tard, la performance des estimateurs dépend crucialement des  $h_t$ , c'est-à-dire des éléments diagonaux de la matrice  $\mathbf{P}_X$ ; les performances des tests basés sur toutes la HCCME en échantillon fini seront d'autant plus faibles que les  $h_t$  les plus élevés seront grands. Quand la matrice  $\mathbf{X}$  est générée comme l'ont fait MacKinnon et White, avec  $n = 50c$ , tous les  $h_t$  doivent approcher zéro à un taux proportionnel à  $1/c$  (et ensuite aussi à  $1/n$ ). Ainsi MacKinnon et White étaient assurés de trouver une amélioration rapide des résultats au fur et à mesure que la

taille de l'échantillon augmentait. Par contraste, Cragg (1983), en réalisant des expériences Monte Carlo sur un problème connexe (consulter la Section 17.3), a généré les  $\mathbf{X}_t$  de façon aléatoire à partir de la distribution lognormale. Cette distribution possède une longue queue de droite et génère ainsi de temps à autres des valeurs élevées pour quelques  $\mathbf{X}_t$ . Celles-ci produisent des valeurs relativement grandes de  $h_t$ , et il en résulte que les valeurs les plus grandes de  $h_t$  tendent vers zéro à un taux beaucoup plus faible que  $1/n$ . Ainsi, comme l'analyse de Chesher-Jewitt l'aurait prédit, Cragg a trouvé que la performance en échantillon fini n'a été améliorée que très légèrement quand la taille de l'échantillon avait augmenté.

Plus récemment, Chesher et Peters (1994) ont montré que les distributions de nombreux estimateurs qui intéressent les économètres dépendent crucialement de la manière dont les régresseurs sont distribués. Si les régresseurs sont distribués symétriquement par rapport à leurs médianes, ces estimateurs auront des propriétés particulières qui ne sont pas valables en général. Puisque les régresseurs utilisés dans les expériences Monte Carlo pourraient bien être symétriquement distribués, il existe un risque que les résultats de telles expériences puissent être sérieusement trompeurs.

Les exemples précédents devraient faciliter la compréhension de deux éléments. Tout d'abord, la manière dont les  $\mathbf{X}_t$  sont générés peut compter. Les chercheurs devraient donc toujours réfléchir avec soin à la façon de générer leurs  $\mathbf{X}_t$ . En second lieu, une bonne compréhension théorique d'un problème peut rendre les expériences Monte Carlo plus informatives et empêcher des conclusions erronées qui peuvent provenir d'aspects apparemment mineurs de la conception expérimentale.

Une des phases les plus ardues de n'importe quelle expérience Monte Carlo consiste à présenter les résultats. Cette phase est souvent beaucoup plus difficile qu'elle ne paraît. Nous discutons ici brièvement de ces problèmes. Une méthode parfois très utile, à savoir l'estimation des surfaces de réponse, ne sera pas traitée ici mais sera largement discutée dans la Section 21.7.

Quand on présente les résultats sous forme de tableau, il est facile de noyer le lecteur. En particulier si plusieurs estimateurs ou statistiques de test doivent être comparés, il est important de rendre les comparaisons aussi lisibles que possible. Par exemple, si l'on est intéressé par l'erreur quadratique moyenne (MSE) de plusieurs estimateurs en compétition, il pourrait être bien plus intéressant de présenter les résultats sous forme de ratios relatifs à un cas de référence, plutôt que de présenter simplement les résultats pour chaque estimateur séparément. Un estimateur relativement simple et bien connu pourrait servir de référence, et les résultats de chacun des autres estimateurs pourraient alors être présentés comme le ratio de la MSE de cet estimateur par la MSE de l'estimateur de référence. Un tel tableau serait très lisible parce que des nombres inférieurs à 1 indiqueraient une meilleure performance que celle de la référence, tandis que des nombres supérieurs à 1 indiqueraient des performances plus faibles. Pour éviter de présenter un grand nombre

d'écarts types expérimentaux, ces ratios pourraient être marqués (en utilisant des symboles tels que \*, †, ou \*\*) pour indiquer s'ils diffèrent de l'unité de manière significative.

Les expérimentateurs présentent souvent simplement des tableaux de moyennes estimées, de variances, et peut-être de coefficients d'asymétrie et d'aplatissement pour plusieurs estimateurs ou statistiques de test différents. Dans le cas des statistiques de test, les probabilités d'aire de queue, c'est-à-dire les niveaux estimés, sont souvent également présentées. De tels tableaux ne sont pas toujours très lisibles. Les méthodes graphiques de présentation peuvent parfois être des alternatives très précieuses, bien qu'elles doivent être utilisées avec modération en fonction de l'espace disponible. Dans le cas des statistiques de test en compétition, on pourrait tracer des courbes de niveau-puissance empiriques (consulter la Section 12.2) de plusieurs statistiques de test sur les mêmes axes. Ceci montrera clairement si une quelconque statistique de test a substantiellement un pouvoir plus ou moins fort que les autres pour un niveau donné; Davidson et MacKinnon (1982) fournissent un exemple. Dans le cas d'estimateurs en compétition, on peut simplement dessiner les fonctions de distribution empiriques de tous les estimateurs sur les mêmes axes, comme dans les Figures 7.1, 7.2, et 18.1. Les différences qualitatives majeures entre les estimateurs en compétition devraient alors être très claires. En outre, étant facile à comprendre, cette approche simplifie le traitement des estimateurs qui manquent de moments (tels que LIML). Pour ces estimateurs, les MSE peuvent bien entendu être extrêmement trompeuses; consulter Sargan (1982).

## 21.5 RÉDUCTION DE VARIANCE: VARIABLES ANTITHÉTIQUES

Comme nous l'avons vu, l'obtention de résultats suffisamment précis à partir d'une expérience Monte Carlo peut parfois nécessiter le calcul d'un grand nombre de répétitions. Ceci n'est pas toujours réalisable. Dans certains cas, le nombre de répétitions nécessaire peut être réduit de manière significative en utilisant certaines techniques de réduction de variance des résultats expérimentaux. Dans la littérature économétrique, les techniques de réduction de variance principalement étudiées sont l'utilisation des **variables antithétiques** et des **variables de contrôle**. Nous discutons de la première méthode dans cette section et de la suivante dans la prochaine section.

L'idée des variables antithétiques consiste à calculer deux estimations différentes de la quantité d'intérêt de telle manière que les deux estimations soient corrélées négativement. Leur moyenne sera ensuite substantiellement plus précise que chacune d'elles prises individuellement. Supposons que l'on veuille estimer une quantité quelconque  $\theta$ , et que dans une seule expérience Monte Carlo nous puissions obtenir deux estimateurs sans biais de  $\theta$ , disons  $\hat{\theta}$  et  $\hat{\theta}$ . Ces deux estimateurs sont les variables antithétiques. Ensuite

l'estimateur pondéré

$$\bar{\theta} = \frac{1}{2}(\acute{\theta} + \grave{\theta}) \quad (21.07)$$

a la variance

$$V(\bar{\theta}) = \frac{1}{4}(V(\acute{\theta}) + V(\grave{\theta}) + 2\text{Cov}(\acute{\theta}, \grave{\theta})),$$

où  $V(\acute{\theta})$  et  $V(\grave{\theta})$  désignent les variances de  $\acute{\theta}$  et  $\grave{\theta}$ . Si  $\text{Cov}(\acute{\theta}, \grave{\theta})$  est négative,  $V(\bar{\theta})$  sera plus petite que  $\frac{1}{4}(V(\acute{\theta}) + V(\grave{\theta}))$ , qui est la variance que nous aurions obtenue avec le même nombre de répétitions pour estimer  $\theta$  à partir de deux expériences indépendantes. Ainsi l'intensité de l'avantage que nous pouvons retirer en utilisant des variables antithétiques dépend de l'intensité de la corrélation négative entre  $\acute{\theta}$  et  $\grave{\theta}$ .

Nous pourrions nous demander pourquoi  $\acute{\theta}$  et  $\grave{\theta}$  doivent avoir la même pondération dans le calcul de  $\bar{\theta}$ . Considérons alors l'estimateur pondéré

$$\ddot{\theta} \equiv w\acute{\theta} + (1-w)\grave{\theta}.$$

Si nous annulons la dérivée de la variance de  $\ddot{\theta}$  par rapport à  $w$ , nous avons

$$w = \frac{V(\grave{\theta}) - \text{Cov}(\acute{\theta}, \grave{\theta})}{V(\acute{\theta}) + V(\grave{\theta}) - 2\text{Cov}(\acute{\theta}, \grave{\theta})},$$

qui est satisfaite en posant  $w = \frac{1}{2}$  lorsque  $V(\acute{\theta}) = V(\grave{\theta})$ . Dans la plupart des cas, les variances des deux estimateurs seront égales, de sorte que leur attribuer un poids égal sera optimal.

Une manière d'implémenter la méthode des variables antithétiques dans le cas des modèles de régression consiste à utiliser chaque ensemble d'aléas généré deux fois, avec le signe opposé la seconde fois. Supposons, par exemple, que nous désirions estimer la moyenne de l'estimation NLS  $\hat{\alpha}$  de l'exposant dans le modèle de régression non linéaire

$$y_t = \beta X_t^\alpha + u_t. \quad (21.08)$$

Pour chaque ensemble d'aléas  $\mathbf{u}^j$ , nous pourrions générer deux réalisations de  $\mathbf{y}$ , avec les  $i^{\text{ième}}$  éléments

$$\acute{y}_t^j = \beta X_t^\alpha + u_t^j \quad \text{et} \quad \grave{y}_t^j = \beta X_t^\alpha - u_t^j.$$

Nous pourrions alors estimer le modèle en utilisant chacun de ces deux ensembles de données, générer ainsi deux estimations différentes de  $\alpha$ ,  $\acute{\alpha}_j$  et  $\grave{\alpha}_j$ . Après  $N$  doubles répétitions, nous pourrions alors construire l'estimateur

$$\bar{\alpha} = \frac{1}{2N} \sum_{j=1}^N (\acute{\alpha}_j + \grave{\alpha}_j),$$

qui est l'analogue de l'estimateur pondéré (21.07). La variance de  $\bar{\alpha}$  pourrait alors être estimée par

$$\frac{1}{N(N-1)} \sum_{j=1}^N \left( \frac{1}{2}(\alpha_j + \dot{\alpha}_j) - \bar{\alpha} \right)^2. \quad (21.09)$$

Puisque  $\bar{\alpha}$  est une simple moyenne de  $\bar{\alpha}_j \equiv \frac{1}{2}(\alpha_j + \dot{\alpha}_j)$  pour  $j = 1, \dots, N$ , (21.09) est simplement l'estimation ordinaire de la variance d'une moyenne d'échantillon.

Il est clair que cette méthode fonctionnera extrêmement bien dans le cas des modèles de régression linéaires à régresseurs fixes. Pour le modèle  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , la  $j^{\text{ième}}$  double répétition donnerait

$$\begin{aligned} \hat{\boldsymbol{\beta}}^j &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{\mathbf{y}}^j = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}^j) \quad \text{et} \\ \dot{\boldsymbol{\beta}}^j &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \dot{\hat{\mathbf{y}}}^j = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{u}^j). \end{aligned}$$

Par conséquent, nous voyons que

$$\begin{aligned} \bar{\boldsymbol{\beta}} &\equiv \frac{1}{2}(\hat{\boldsymbol{\beta}}^j + \dot{\boldsymbol{\beta}}^j) \\ &= \frac{1}{2}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0 + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}^j - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}^j) = \boldsymbol{\beta}_0. \end{aligned}$$

Ainsi, dans une seule double répétition, nous pourrions obtenir une réponse sans erreur expérimentale. Ceci survient parce que  $\hat{\boldsymbol{\beta}}^j$  et  $\dot{\boldsymbol{\beta}}^j$  sont parfaitement corrélés négativement.

La corrélation négative parfaite des variables ne se produira pas en général. Quand elle survient, le problème est habituellement tellement simple qu'il n'est pas nécessaire d'exécuter des expériences Monte Carlo (bien que parfois une très petite expérience Monte Carlo, qui consiste juste en une double répétition utilisant des variables antithétiques, puisse nous révéler qu'un estimateur est sans biais plus facilement que ne le ferait une analyse théorique). Cependant, une corrélation négative moins parfaite survient souvent, et elle signifie que dans certains cas l'utilisation de variables antithétiques peut grandement réduire le nombre de répétitions nécessaires pour estimer les premiers moments d'un estimateur. Hendry et Trivedi (1972) ont utilisé la technique pour étudier les estimateurs de certains modèles dynamiques, et Mikhail (1972, 1975) l'a utilisée pour étudier certains estimateurs d'équations simultanées.

Considérons à nouveau l'exemple (21.08). Nous avons mené une petite expérience Monte Carlo basée sur cet exemple, avec un échantillon 50 observations, et un seul ensemble de  $\mathbf{X}_t$  généré à partir de la distribution uniforme sur l'intervalle (5, 15) et les paramètres  $\alpha_0 = 0.5$ ,  $\beta_0 = 1.0$ , et  $\sigma_0^2 = 1.0$  (ici  $\sigma_0^2$

**Tableau 21.1** Moyennes et Ecart Types des Estimations Monte Carlo

$\hat{\alpha}$ :	0.515960	(0.006709)	$\hat{\beta}$ :	1.019957	(0.016002)
$\hat{\alpha}$ :	0.488785	(0.006627)	$\hat{\beta}$ :	1.088944	(0.016998)
$\bar{\alpha}$ :	0.502372	(0.000425)	$\bar{\beta}$ :	1.054451	(0.003404)

est la variance des  $u_t$ , supposés normaux). Les résultats issus de 500 doubles répétitions sont rapportés dans le Tableau 21.1.

Dans ce cas, les gains provenant de l'usage des variables antithétiques sont apparemment très importants. L'écart type de  $\bar{\alpha}$  est 15.7 fois plus petit que la moyenne des écarts types de  $\hat{\alpha}$  et  $\hat{\alpha}$ . Ceci signifie que  $\bar{\alpha}$ , qui est basé sur 1000 répétitions, est aussi précis que l'estimation naïve Monte Carlo basée sur approximativement 246,000 répétitions! Les gains sont moins flagrants dans le cas de  $\beta$ , mais ils sont encore très conséquents. L'écart type de  $\bar{\beta}$  est 4.8 fois plus petit que la moyenne des écarts types de  $\hat{\beta}$  et  $\hat{\beta}$ , ce qui signifie qu'il est aussi précis qu'une estimation naïve basée sur environ 23,500 répétitions. Du fait de la précision de  $\bar{\alpha}$  et  $\bar{\beta}$ , nous pouvons voir que les NLS produisent des estimations légèrement biaisées dans ce cas: les  $t$  de Student pour les hypothèses nulles que les moyennes des estimations de  $\alpha$  et  $\beta$  sont les véritables valeurs 0.5 et 1.0 sont, respectivement, 5.58 et 16.00.

Bien que des variables antithétiques du type de celles décrites puissent réellement réduire le nombre de répétitions Monte Carlo nécessaires pour préciser les estimations des *moyennes* des estimateurs, il n'existe aucune aide possible pour estimer de nombreuses autres caractéristiques de leurs distributions. Par exemple, dans le cas OLS discuté au préalable, la matrice de covariance estimée des  $\hat{\beta}^j$  est

$$\frac{1}{N} \sum_{j=1}^N (\hat{\beta}^j - \beta_0) (\hat{\beta}^j - \beta_0)^\top,$$

et la matrice de covariance estimée des  $\hat{\beta}^j$  est

$$\frac{1}{N} \sum_{j=1}^N (\hat{\beta}^j - \beta_0) (\hat{\beta}^j - \beta_0)^\top.$$

Il est facile de voir que

$$\begin{aligned} (\hat{\beta}^j - \beta_0) (\hat{\beta}^j - \beta_0)^\top &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} \mathbf{u}^j (\mathbf{u}^j)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} (-\mathbf{u}^j) (-\mathbf{u}^j)^\top (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\hat{\beta}^j - \beta_0) (\hat{\beta}^j - \beta_0)^\top. \end{aligned}$$

Ainsi les matrices de covariance estimées des deux variables antithétiques seront identiques. Alors, du point de vue de l'estimation de la matrice de covariance de l'estimateur, la seconde variable antithétique ne fournit aucune information utile. Dans une situation réaliste, les matrices de covariance des deux variables antithétiques ne seront jamais corrélées à la perfection, mais pourront être corrélées positivement. L'estimation antithétique de la matrice de covariance sera par conséquent moins efficace que l'estimation naïve basée sur le même nombre de répétitions.

## 21.6 RÉDUCTION DE VARIANCE: VARIABLES DE CONTRÔLE

La seconde technique largement utilisée pour la réduction de variance consiste à employer des variables de contrôle. Une **variable de contrôle** est une variable aléatoire dont la distribution (ou du moins certaines propriétés de la distribution) est connue et corrélée avec l'(es) estimateur(s) ou la(les) statistique(s) de test étudiés. La première propriété qu'une variable de contrôle doit posséder est une moyenne de population connue. La divergence entre la moyenne d'échantillon de la variable de contrôle dans l'expérience et sa moyenne de population connue est ensuite utilisée pour améliorer les estimations de l'expérience Monte Carlo. Ceci fonctionne évidemment mieux si la variable de contrôle est fortement corrélée aux estimateurs ou aux statistiques de test de l'expérience concernée.

Typiquement, les variables de contrôle sont des statistiques qui ne pourraient jamais être calculées dans la pratique mais qui peuvent l'être dans le cadre d'une expérience Monte Carlo, parce que le DGP est connu. Par exemple, supposons que l'expérience concerne les estimations de  $\beta$  à partir d'un modèle de régression non linéaire à aléas normaux,

$$\mathbf{y} = \mathbf{x}(\beta) + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}),$$

où  $\mathbf{x}(\beta)$  ne dépend que de  $\beta$  et des régresseurs fixes ou du moins indépendants de  $\mathbf{u}$ . Nous avons vu dans la Section 5.4 que

$$n^{1/2}(\hat{\beta} - \beta_0) = (n^{-1} \mathbf{X}_0^\top \mathbf{X}_0)^{-1} n^{-1/2} \mathbf{X}_0^\top \mathbf{u} + o(1).$$

Ainsi il est naturel de considérer l'utilisation du vecteur

$$\check{\beta} = (\mathbf{X}_0^\top \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{u}$$

comme une source de variables de contrôle. Ce vecteur sera bien évidemment normal avec un vecteur d'espérances nulles et une matrice de covariance  $\sigma_0^2 (\mathbf{X}_0^\top \mathbf{X}_0)^{-1}$ . Il serait impossible de calculer  $\check{\beta}$  à partir d'un ensemble de données réelles, mais dans le cadre d'une expérience Monte Carlo, cela est parfaitement réalisable. Nous connaissons  $\beta_0$  et par conséquent  $\mathbf{X}_0 \equiv \mathbf{X}(\beta_0)$ .

Ces vecteur et matrice connus, et avec le vecteur d'erreur  $\mathbf{u}^j$  généré à chaque répétition, nous pouvons facilement calculer  $\hat{\beta}^j$ .

Supposons que  $\theta \equiv \theta(\hat{\beta})$  soit une quantité scalaire quelconque dont nous désirons calculer la moyenne en utilisant les résultats de l'expérience Monte Carlo. Par exemple, si nous étions intéressés par le biais de  $\hat{\beta}_2$ ,  $\theta$  serait  $\hat{\beta}_2 - \beta_{20}$ ; si nous étions intéressés par l'erreur quadratique moyenne de  $\hat{\beta}_3$ ,  $\theta$  serait  $(\hat{\beta}_3 - \beta_{30})^2$ ; si nous étions intéressés par le niveau d'un test,  $\theta$  serait 1 si le test rejetait l'hypothèse et 0 sinon; et ainsi de suite. A chaque répétition, nous obtenons  $t_j$ , une réalisation de  $\theta$ , égale à  $\theta(\hat{\beta}^j)$ . Nous obtenons également une variable de contrôle  $\tau_j$ , qui serait normalement une certaine fonction de  $\hat{\beta}$ . Les  $\tau_j$  doivent avoir une moyenne nulle et une variance finie, qui peut être inconnue. Si nous sommes intéressés par le biais de  $\hat{\beta}_2$ , par exemple, le choix naturel pour  $\tau$  serait  $\hat{\beta}_2 - \beta_{20}$ . Cependant, dans certains autres cas, il n'est pas évident de savoir comment choisir  $\tau$ , et il peut exister plusieurs choix possibles.

Si la variable de contrôle  $\tau$  n'était pas disponible, nous estimerions  $\theta$  par

$$\bar{\theta} \equiv \frac{1}{N} \sum_{j=1}^N t_j,$$

et cet estimateur naïf aurait une variance  $V(\bar{\theta}) = N^{-1}V(t)$ , qui pourrait être estimée par

$$\hat{V}(\bar{\theta}) = \frac{1}{N(N-1)} \sum_{j=1}^N (t_j - \bar{\theta})^2.$$

Quand la variable de contrôle  $\tau$  est disponible,  $\bar{\theta}$  ne sera plus optimale dans la plupart des cas. Considérons alors l'estimateur de la variable de contrôle (CV)

$$\check{\theta}(\lambda) \equiv \bar{\theta} - \lambda \bar{\tau}, \quad (21.10)$$

où  $\bar{\tau}$  est la moyenne d'échantillon des  $\tau_j$ . Cet estimateur implique de soustraire à  $\bar{\theta}$  un certain multiple  $\lambda$  de la moyenne d'échantillon des variables de contrôle; le choix de  $\lambda$  sera discuté dans le prochain paragraphe. En moyenne, ce qui est soustrait sera nul, puisque  $\tau_j$  a une moyenne de population nulle. Ceci implique que  $\check{\theta}(\lambda)$  doit avoir la même moyenne de population que  $\bar{\theta}$ . Mais, dans n'importe quel échantillon donné, la moyenne des  $\tau_j$  sera non nulle. Si, par exemple, elle est positive, et si  $\tau_j$  et  $t_j$  sont fortement corrélés positivement, il est très probable que  $\bar{\theta}$  excédera également sa moyenne de population. Ainsi, en soustrayant à  $\bar{\theta}$  un multiple de la moyenne des  $\tau_j$ , nous aurons de fortes chances d'obtenir une meilleure estimation de  $\theta$ .

La variance de l'estimateur CV (21.10) est

$$V(\check{\theta}(\lambda)) = V(\bar{\theta}) + \lambda^2 V(\bar{\tau}) - 2\lambda \text{Cov}(\bar{\theta}, \bar{\tau}). \quad (21.11)$$



Il est facile de minimiser cette expression par rapport à  $\lambda$ . La valeur optimale de  $\lambda$  se trouve être

$$\lambda^* = \frac{\text{Cov}(\bar{\theta}, \bar{\tau})}{V(\bar{\tau})}. \quad (21.12)$$

En substituant (21.12) dans (21.11), la variance de  $\ddot{\theta}(\lambda^*)$  est

$$V(\ddot{\theta}(\lambda^*)) = V(\bar{\theta}) - \frac{\text{Cov}(\bar{\theta}, \bar{\tau})^2}{V(\bar{\tau})} = (1 - \rho^2)V(\bar{\theta}), \quad (21.13)$$

où

$$\rho \equiv \frac{\text{Cov}(\bar{\theta}, \bar{\tau})}{(V(\bar{\tau})V(\bar{\theta}))^{1/2}}$$

est la corrélation entre les  $t_j$  et les  $\tau_j$ . À partir de (21.13), il est clair qu'à chaque fois que cette corrélation n'est pas nulle, il y aura un certain avantage à utiliser la variable de contrôle. Si la corrélation est forte, l'avantage peut être très important. Par exemple, si  $\rho = 0.95$ , la variance de  $\ddot{\theta}(\lambda^*)$  sera 0.0975 fois la variance de  $\bar{\theta}$ . L'utilisation de la variable de contrôle sera alors équivalente à accroître le nombre de répétitions par un facteur de 10.26.

Quand la taille d'échantillon  $n$  augmente, la corrélation entre la variable de contrôle et la quantité d'intérêt devrait augmenter, parce que la distribution en échantillon fini de cette dernière devrait s'approcher de sa distribution asymptotique quand  $n$  augmente. Par conséquent, le gain d'efficacité provenant de l'utilisation de la variable de contrôle devrait être d'autant plus important que  $n$  est grand. Ceci est commode parce que le coût de réalisation des expériences Monte Carlo est souvent presque proportionnel à  $nN$ , et l'efficacité croissante de l'estimation quand  $n$  augmente permettra de réduire  $N$  dans le même temps.

Même si  $V(\bar{\tau})$  sera souvent connue,  $\text{Cov}(\bar{\theta}, \bar{\tau})$  ne le sera presque jamais. Ainsi, nous aurons généralement à estimer  $\lambda^*$  d'une manière quelconque. Une littérature fournie sur les méthodes Monte Carlo — par exemple, Hammersley et Handscomb (1964) et Hendry (1984) — ne cherche pas à utiliser  $\lambda^*$  mais pose au contraire  $\lambda = 1$ . À partir de (21.12) et de la définition de  $\rho$ , nous voyons que

$$\lambda^* = \rho \left( \frac{V(\bar{\theta})}{V(\bar{\tau})} \right)^{1/2}.$$

Ceci implique que  $\lambda = 1$  sera un bon choix si  $\rho$  est proche de 1 et  $V(\bar{\theta})$  proche de  $V(\bar{\tau})$ , mais ce choix n'est pas le meilleur en général. Dans de nombreux cas,  $\rho$  peut être significativement inférieur à 1 mais encore suffisamment grand pour rendre intéressante l'utilisation des variables de contrôle, et dans d'autres cas  $V(\bar{\tau})$  peut ne pas être proche de  $V(\bar{\theta})$  quand on utilise la définition la plus naturelle de  $\tau$ . Ainsi, nous préférons, en général, estimer  $\lambda^*$ . La manière la plus facile d'y parvenir est d'exécuter la régression

$$t_j = \theta + \lambda \tau_j + \text{résidu}. \quad (21.14)$$

Comme la notation le suggère, cette régression ne fournit pas seulement une estimation de  $\lambda^*$  mais également une estimation de  $\theta$ . Cette dernière est en fait asymptotiquement équivalente à  $\ddot{\theta}(\lambda^*)$ . Ainsi, comme nous allons maintenant le montrer, la régression (21.14) fournit une manière remarquablement simple de calculer un estimateur CV asymptotiquement optimal.

L'estimation OLS de  $\lambda$  à partir de (21.14) est

$$\hat{\lambda} = (\boldsymbol{\tau}^\top \mathbf{M}_\boldsymbol{\iota} \boldsymbol{\tau})^{-1} \boldsymbol{\tau}^\top \mathbf{M}_\boldsymbol{\iota} \mathbf{t},$$

où  $\mathbf{t}$ ,  $\boldsymbol{\tau}$ , et  $\boldsymbol{\iota}$  sont des vecteurs d'éléments types  $t_j$ ,  $\tau_j$ , et 1, et  $\mathbf{M}_\boldsymbol{\iota}$  est la matrice  $\mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}^\top \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top$  qui calcule les écart types provenant de la moyenne. Il est facile de voir que  $\hat{\lambda}$  est juste la covariance d'échantillon de  $\mathbf{t}$  et  $\boldsymbol{\tau}$ , divisée par la variance d'échantillon de  $\boldsymbol{\tau}$ . C'est donc la contrepartie empirique de  $\lambda^*$ . Comme les résidus d'une régression linéaire avec un terme constant doivent avoir une somme nulle, l'estimation OLS de  $\theta$  peut être écrite comme

$$\hat{\theta} = \bar{\theta} - \hat{\lambda} \bar{\tau}.$$

Ceci montre clairement que l'estimation OLS  $\hat{\theta}$  est égale à  $\ddot{\theta}(\hat{\lambda})$ . Puisque  $\hat{\lambda}$  converge vers  $\lambda^*$  sous des hypothèses plutôt faibles,  $\hat{\theta}$  sera asymptotiquement équivalente à  $\ddot{\theta}(\lambda^*)$ .

L'exécution de la régression (21.14) ne fournit pas seulement l'estimation CV  $\hat{\theta}$  mais aussi une estimation de la variance de cette estimation, dont nous avons besoin pour calibrer la précision des résultats et décider si  $N$  est suffisamment grand. Cette variance estimée est

$$\hat{\sigma}^2 (\boldsymbol{\iota}^\top \mathbf{M}_\boldsymbol{\tau} \boldsymbol{\iota})^{-1},$$

où  $\hat{\sigma}$  est l'écart type de la régression (21.14). Ici, le second facteur doit tendre vers  $N^{-1}$ , puisque  $\boldsymbol{\tau}$  (parce qu'il a une moyenne nulle) n'a pas asymptotiquement de pouvoir explicatif sur  $\boldsymbol{\iota}$ . Par conséquent,  $N^{-1} \hat{\sigma}^2$  serait aussi une estimation valable de la variance de  $\hat{\theta}$ . Puisque  $\sigma^2$  est la variance de la partie des  $t_j$  qui ne peut être expliquée par les  $\tau_j$ , il est clair que la précision de l'estimation CV  $\hat{\theta}$  sera d'autant meilleure que l'ajustement de la régression (21.14) sera bon.

Une fois énoncé le problème en termes de la régression (21.14), il devient clair que le lien entre  $\theta$  et les  $\tau_j$  n'est pas forcément étroit. N'importe quelle variable aléatoire qui peut être calculée avec  $t_j$  peut être utilisée comme variable de contrôle pourvu qu'elle soit corrélée à  $t_j$  (soit positivement, soit négativement) et ait une moyenne nulle, une variance finie, et une covariance finie avec  $t_j$ . Puisque c'est le cas, il peut exister plus d'un choix naturel pour  $\boldsymbol{\tau}$  dans de nombreuses situations. Heureusement, la formulation du problème en régression linéaire rend évidente la manière de traiter des variables de contrôle multiples. La généralisation appropriée de (21.14) est

$$\mathbf{t} = \theta \boldsymbol{\iota} + \mathbf{T} \boldsymbol{\lambda} + \text{résidus}, \quad (21.15)$$

où  $\mathbf{T}$  est une matrice de dimension  $N \times c$ , dont chaque colonne se compose des observations sur une des  $c$  variables de contrôle. Puisque toutes les colonnes de  $\mathbf{T}$  ont une moyenne nulle, l'estimation OLS de  $\theta$  à partir de cette régression fournira encore une fois l'estimation que nous cherchons.<sup>2</sup> Cette estimation est

$$\hat{\theta} = (\boldsymbol{\iota}^\top \mathbf{M}_T \boldsymbol{\iota})^{-1} \boldsymbol{\iota}^\top \mathbf{M}_T \mathbf{t},$$

où  $\mathbf{M}_T = \mathbf{I} - \mathbf{T}(\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{T}^\top$ . Puisque  $N^{-1} \boldsymbol{\iota}^\top \mathbf{M}_T \boldsymbol{\iota}$  tend vers l'unité quand  $N$  tend vers l'infini, il est facile de voir que la variance de  $\hat{\theta}$  est encore une fois  $N^{-1} \sigma^2$ , où  $\sigma$  est le véritable écart type de la régression (21.15). Ainsi, notre objectif dans le choix des variables de contrôle consiste à rendre l'ajustement de la régression (21.15) aussi bon que possible.

Supposons que nous soyons intéressés par le niveau  $p$  d'un test quelconque, qui correspond à la probabilité que le test rejettera l'hypothèse nulle quand elle est vraie. Nous obtenons  $N$  observations  $T_j$  sur la statistique de test et  $N$  observations sur une variable de contrôle  $C_j$  de distribution connue. Construisons une variable 0-1  $t_j$  de telle sorte que  $t_j = 1$  si  $T_j$  excède une certaine valeur critique et que  $t_j = 0$  sinon. Alors la moyenne des  $t_j$  est une estimation naïve de  $p$ . Davidson et MacKinnon (1981b) et Rothery (1982) ont considéré ce problème en détail et proposé une méthode d'utilisation de la variable de contrôle pour estimer  $p$  basée sur la méthode du maximum de vraisemblance. Il en ressort que leur estimateur est identique à celui de l'estimateur OLS de  $\theta$  issu de la régression (21.14), où  $\tau_j$  est une variable égale à  $1 - s$  quand  $C_j$  excède la valeur critique pour un test de niveau  $s$ , et  $-s$  sinon. Puisque la probabilité que  $C_j$  excédera la valeur critique est  $s$ ,  $\tau_j$  définie de cette manière a manifestement une moyenne de population nulle. Cette technique nécessite un choix de  $s$ . Comme nous désirons maximiser la corrélation entre les  $t_j$  et les  $\tau_j$ , il semble logique d'assimiler  $s$  au nombre de rejets réellement observés avec  $T_j$ . Quoi qu'il en soit, le choix des valeurs critiques est forcément arbitraire.

Laisser  $\tau_j$  prendre seulement deux valeurs ne peut pas être optimal, puisque nous perdons une certaine information dans les  $C_j$ . On pourrait tout aussi simplement utiliser n'importe quelle fonction de  $C_j$  moins sa moyenne pour  $\tau_j$ , fonction de nous savons fortement corrélée à  $t_j$ . Vue l'étendue des possibilités, il semblerait naturel d'utiliser plus d'une d'entre elles. Par exemple, si nous savons que  $C_j$  est distribuée suivant la  $N(0,1)$ , et sommes intéressés par un test bilatéral, on pourrait utiliser  $C_j^2 - 1$  comme variable de contrôle. Elle sera d'espérance nulle, puisque l'espérance d'une variable aléatoire du  $\chi^2(1)$  est 1, et elle devrait être corrélée à  $t_j$ . On pourrait tout

<sup>2</sup> Il est intéressant d'observer que la régression (21.15) est formellement la même que la régression (16.63), la version de Tauchen (1985) de la régression de test OPG. Les deux régressions fournissent une manière d'estimer efficacement la moyenne de la régressande en tenant compte de la corrélation entre elle et les autres régresseurs, asymptotiquement orthogonaux au terme constant.

aussi bien l'utiliser avec une ou plusieurs variables de contrôle binaires du type décrit précédemment. L'expérience suggère que l'utilisation de plusieurs variables de contrôle produit généralement une estimation plus précise de  $\theta$  que lorsqu'il n'y en a qu'une seule. Dans la pratique, il est facile d'expérimenter des variables de contrôle diverses en examinant celles qui sont significatives dans la régression (21.15).

L'emploi des régressions (21.14) et (21.15) a été préconisé pendant un certain temps dans la littérature de recherche opérationnelle; consulter Lavenberg et Welch (1981) et Ripley (1987). Ces procédures ont été exposées et développées dans Davidson et MacKinnon (1993), lesquels ont montré comment les utiliser pour l'estimation des quantiles aussi bien que pour l'estimation des moments et des aires de queues, ainsi que la façon de construire les  $\tau$  approximativement optimaux dans plusieurs cas d'intérêt. En particulier, pour l'estimation des niveaux et des puissances de test, une manière fut proposée pour construire des variables de contrôle plus intelligemment, mais plus difficilement, qu'avec la manière à deux valeurs discutée précédemment.

Pour illustrer l'usage des variables de contrôle, nous considérerons un exemple simple discuté par Hendry (1984). Il s'agit du modèle AR(1) stationnaire à aléas normaux:

$$y_t = \beta y_{t-1} + u_t, \quad u_t \sim N(0, \sigma^2), \quad t = 1, \dots, n. \quad (21.16)$$

Nous supposons que  $|\beta| < 1$ , qui correspond à la condition de stationnarité, et que  $y_0 = 0$ . La stationnarité implique que  $y_t \sim N(0, \sigma^2/(1 - \beta^2))$ . Supposons que nous soyons intéressés par la moyenne de  $\hat{\beta}$ , l'estimation OLS de  $\beta$ . Il est facile de voir qu'à la fois la valeur de  $\hat{\beta}$  et sa distribution de probabilité sont invariantes à la valeur de  $\sigma$  dans le DGP, disons  $\sigma_0$ , mais que ses propriétés peuvent bien dépendre à la fois de  $\beta_0$  et de la taille d'échantillon  $n$ . Une recherche sérieuse s'attacherait par conséquent à déterminer le type de dépendance de la moyenne de  $\hat{\beta}$  à  $\beta_0$  et  $n$ ; consulter la Section 21.7 qui suit. Puisque nous sommes ici beaucoup intéressés par l'illustration de l'utilisation des variables de contrôle, nous ne considérerons que quelques cas particuliers.<sup>3</sup>

L'estimation OLS  $\hat{\beta}$ , en supposant  $y_0$  connue, est

$$\hat{\beta} = \frac{\sum_{t=1}^n y_t y_{t-1}}{\sum_{t=1}^n y_{t-1}^2}.$$

Sous le DGP caractérisé par  $\beta_0$ , ceci devient

$$\frac{\sum_{t=1}^n (\beta_0 y_{t-1} + u_t) y_{t-1}}{\sum_{t=1}^n y_{t-1}^2} = \beta_0 + \frac{\sum_{t=1}^n u_t y_{t-1}}{\sum_{t=1}^n y_{t-1}^2}. \quad (21.17)$$

<sup>3</sup> Notons que, bien que (21.16) ressemble à un modèle de régression, des variables antithétiques ne sont pas utiles ici. Si l'on génère deux ensembles de données avec des vecteurs de perturbations  $\mathbf{u}$  et  $-\mathbf{u}$ , les estimations de  $\beta$  obtenues sont identiques.

**Tableau 21.2** Estimations CV et Naïves de la Moyenne de  $\hat{\beta}$ 

$\beta_0$	$n$	Naïve	$\hat{\lambda}$	CV Optimale
0.1	25	0.091814 (0.001932)	0.927	0.091461 (0.000548)
0.1	100	0.096499 (0.000978)	0.982	0.097889 (0.000140)
0.1	400	0.099731 (0.000502)	0.995	0.099499 (0.000036)
0.5	25	0.465589 (0.001745)	0.934	0.464972 (0.000666)
0.5	100	0.490394 (0.000876)	0.982	0.490013 (0.000182)
0.5	400	0.497774 (0.000439)	0.991	0.497430 (0.000048)
0.9	25	0.843872 (0.001188)	0.958	0.843656 (0.000841)
0.9	100	0.882824 (0.000497)	0.987	0.882975 (0.000246)
0.9	400	0.895824 (0.000228)	0.992	0.895530 (0.000066)

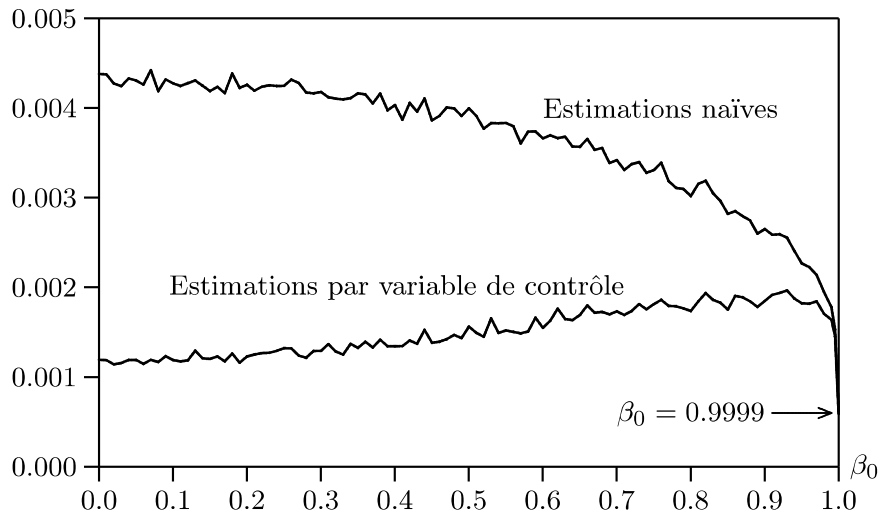
Bien que le numérateur du second terme du membre de droite de (21.17) ait une moyenne nulle, il n'est pas indépendant du dénominateur, et donc  $E(\hat{\beta}) \neq \beta_0$ . Cependant, la théorie asymptotique nous dit que  $\hat{\beta}$  est convergente et asymptotiquement normale, puisque  $n^{1/2}(\hat{\beta} - \beta_0) \stackrel{a}{\sim} N(0, 1 - \beta_0^2)$ .

Considérons maintenant la variable de contrôle

$$\tau = n^{-1/2} \sum_{t=1}^n u_t y_{t-1}, \quad (21.18)$$

qui, à partir de (21.17), est  $n^{-1/2}$  fois le numérateur de la partie stochastique de  $\hat{\beta}$ . La distribution en échantillon fini de la variable de contrôle  $\tau$  définie dans (21.18) n'est pas simple. Cependant, il est facile de voir que  $\tau$  a une moyenne nulle. Pourvu que  $|\beta| < 1$ , il est également facile de vérifier que  $\tau$  a une variance finie  $\sigma_0^4/(1 - \beta_0^2)$ . Ainsi, il est légitime d'utiliser  $\tau$  comme variable de contrôle. A partir de (21.17), il est clair qu'asymptotiquement la corrélation entre  $\tau$  et  $\hat{\beta} - \beta_0$  sera unitaire. Par conséquent, il est vraisemblable qu'il y ait une forte corrélation positive en échantillon fini.

Les résultats des 10,000 répétitions pour trois valeurs de  $\beta_0$  et trois valeurs de  $n$  sont présentés dans le Tableau 21.2. Pour chaque  $\beta_0$  et chaque taille d'échantillon, nous présentons deux estimations de la moyenne de  $\hat{\beta}$ : l'estimation naïve qui n'utilise pas de variable de contrôle, et l'estimation CV optimale basée sur l'équation (21.14). Le tableau donne aussi la valeur de  $\lambda$  implicitement utilisée pour calculer cette dernière quand  $\tau$  est transformé de telle sorte qu'il a la même variance, asymptotiquement, que  $\hat{\beta}$ . Les écarts types estimés apparaissent entre parenthèses. Nous voyons que, comme cela est bien connu, l'estimateur OLS de  $\beta$  est toujours biaisé vers zéro et que le biais décline fortement quand  $n$  augmente. Nous voyons également que l'avantage provenant de l'utilisation de la variable de contrôle varie nettement



**Figure 21.3** Écarts type estimés des biais estimés,  $n = 25$

d'un cas à l'autre. Pour un  $\beta_0$  donné, le gain proportionnel augmente avec  $n$ . Pour un  $n$  donné, il décroît quand  $\beta_0$  approche un. Dans le meilleur des cas ( $n = 400$ ,  $\beta_0 = 0.1$ ) le recours à la variable de contrôle a le même effet que l'augmentation de  $N$  de 10,000 à 1.9 million, tandis que dans le pire des cas ( $n = 25$ ,  $\beta_0 = 0.9$ ) il a l'effet d'une augmentation de  $N$  légèrement en dessous de 20,000. Il est intéressant de noter que les valeurs de  $\hat{\lambda}$  sont toujours assez élevées, devenant très proches de 1 pour  $n = 400$ . Evidemment, il serait un peu plus coûteux de poser  $\lambda = 1$  dans cet exemple.

L'intensité de l'utilité des variables de contrôle dépendra souvent dans la pratique des valeurs paramétriques. Ceci est explicitement illustré dans la Figure 21.3, qui montre les écarts types estimés des estimations des variables naïve et de contrôle de  $\beta$ , pour 101 valeurs de  $\beta_0$  allant de zéro à 0.9999 avec des intervalles de 0.01. Nous avons utilisé 0.9999 comme limite supérieure plutôt que 1.0, parce que les données étaient générées suivant l'hypothèse de stationnarité. Les résultats pour l'intervalle allant de zéro à  $-0.9999$  seraient identiques. Chaque estimation est basée sur 2000 répétitions, les irrégularités évidentes sur la figure traduisent l'erreur expérimentale dans l'estimation des écarts types. Il est très clair à partir de la figure que, pour la plupart des valeurs de  $\beta_0$ , les estimations CV sont beaucoup plus efficaces que les estimations naïves. Cependant, quand  $\beta_0 \rightarrow 1$ , les deux ensembles d'estimations, et en particulier les estimations naïves, deviennent soudainement plus efficaces, et il n'existe virtuellement aucun élément permettant de choisir entre les estimations CV et les estimations naïves pour  $\beta_0 > 0.98$ . Ceci explique pourquoi les variables de contrôle n'ont pas été employées dans les expériences Monte Carlo destinées à déterminer les distributions des statistiques de test de racines unitaires et de cointégration (voir les Sections 20.3 et 20.6).

On pourrait très bien être intéressé par d'autres aspects des estimations OLS de  $\beta$  en plus de leur moyenne. Une possibilité, par exemple, est leur erreur quadratique moyenne. Dans ce cas, l'usage de (21.18) comme variable de contrôle n'est plus naturel, mais il semble plausible d'utiliser

$$\frac{1}{n} \sum_{t=1}^n (u_t y_{t-1})^2 - \frac{\sigma_0^4}{1 - \beta_0^2}, \quad (21.19)$$

puisqu'elle mesure la variance du numérateur de la partie stochastique de  $\hat{\beta}$ . Une autre variable de contrôle possible est

$$\frac{1}{n} \sum_{t=1}^n y_{t-1}^2 - \frac{\sigma_0^2}{1 - \beta_0^2}, \quad (21.20)$$

qui est le dénominateur de la partie stochastique de  $\hat{\beta}$ , moins sa moyenne. L'expression (21.20) n'a pas été mentionnée plus tôt comme variable de contrôle possible parce qu'elle s'est révélée complètement inutile dans la régression de la variable de contrôle pour la moyenne de  $\beta$ , mais il s'avère qu'elle est utile dans ce cas.

Le Tableau 21.3 rapporte des estimations naïves et deux ensembles d'estimations CV de l'erreur quadratique moyenne de  $\hat{\beta}$ , pour un découpage identique à celui du Tableau 21.2. L'usage d'une variable de contrôle unique, (21.19), fournit généralement des estimations plus précises que la non utilisation de variable de contrôle; l'usage de deux variables de contrôle, (21.19) et (21.20), fonctionne toujours mieux que l'usage d'une seule. Cependant, les gains relatifs à l'estimateur naïf sont toujours inférieurs à ceux obtenus lorsque l'on a estimé la moyenne; comparer avec le Tableau 21.1. Cela illustre le résultat général selon lequel les variables de contrôle tendent à être les plus utiles pour l'estimation des moyennes et progressivement de moins en moins utiles pour l'estimation des moments supérieurs; consulter Davidson et MacKinnon (1993).

Etant donnée la forte variabilité des gains découlant de l'usage des variables de contrôle, il peut être judicieux dans les cas où les coûts de calcul sont importants d'adapter le nombre de répétitions  $N$ . On pourrait déterminer au préalable le niveau de précision acceptable pour des quantités diverses à estimer, puis calculer ces quantités pour une valeur initiale relativement faible de  $N$  (peut-être 500), et utiliser ces résultats initiaux pour estimer le nombre de répétitions nécessaires pour obtenir des écarts types suffisamment faibles. Alternativement, on pourrait calculer des écarts types des quantités d'intérêt après quelques centaines de répétitions, en s'arrêtant quand ils sont suffisamment faibles. Dans la pratique, peu d'expériences Monte Carlo ont été conçues de cette manière;  $N$  est généralement fixé préalablement, et la précision des estimations est simplement ce qu'il en ressort.

**Tableau 21.3** Estimations CV et Naïves de la MSE de  $\hat{\beta}$ 

$\beta_0$	$n$	Naïve	Une Vble de Contrôle	Deux Vbles de Contrôle
0.1	25	.03739 ( $.510 \times 10^{-3}$ )	.03720 ( $.317 \times 10^{-3}$ )	.03728 ( $.272 \times 10^{-3}$ )
0.1	100	.00959 ( $.134 \times 10^{-3}$ )	.00973 ( $.468 \times 10^{-4}$ )	.00970 ( $.390 \times 10^{-4}$ )
0.1	400	.00252 ( $.351 \times 10^{-4}$ )	.00247 ( $.650 \times 10^{-5}$ )	.00246 ( $.524 \times 10^{-5}$ )
0.5	25	.03161 ( $.522 \times 10^{-3}$ )	.03171 ( $.454 \times 10^{-3}$ )	.03139 ( $.384 \times 10^{-3}$ )
0.5	100	.00777 ( $.734 \times 10^{-4}$ )	.00768 ( $.696 \times 10^{-4}$ )	.00767 ( $.542 \times 10^{-4}$ )
0.5	400	.00193 ( $.281 \times 10^{-4}$ )	.00187 ( $.976 \times 10^{-5}$ )	.00188 ( $.756 \times 10^{-5}$ )
0.9	25	.01725 ( $.413 \times 10^{-3}$ )	.01725 ( $.413 \times 10^{-3}$ )	.01731 ( $.377 \times 10^{-3}$ )
0.9	100	.00277 ( $.563 \times 10^{-4}$ )	.00276 ( $.548 \times 10^{-4}$ )	.00274 ( $.439 \times 10^{-4}$ )
0.9	400	.00054 ( $.922 \times 10^{-5}$ )	.00053 ( $.748 \times 10^{-5}$ )	.00053 ( $.534 \times 10^{-5}$ )

## 21.7 LES SURFACES DE RÉPONSE

Comme nous l'avons souligné auparavant, l'un des aspects les plus difficiles dans n'importe quelle expérience Monte Carlo est de présenter les résultats de façon lisible. Une approche parfois très utile consiste à estimer une **surface de réponse**. Il s'agit simplement d'un modèle de régression pour lequel chaque observation correspond à une expérience, la variable dépendante est une quantité quelconque estimée dans les expériences, et les variables indépendantes sont des fonctions de différentes valeurs paramétriques choisies par l'expérimentateur, et qui caractérisent chaque expérience. Les surfaces de réponse ont été utilisées par Hendry (1979), Mizon et Hendry (1980), Engle, Hendry, et Trumble (1985), Ericsson (1991), et MacKinnon (1991), parmi d'autres; elles sont longuement discutées dans Hendry (1984). Pour les critiques de cette approche, consulter Maasoumi et Phillips (1982), ainsi que la réponse de Hendry (1982).

Si l'on peut trouver une surface de réponse qui explique de façon adéquate les résultats expérimentaux, cette approche qui synthétise les résultats Monte Carlo mérite d'y prêter attention. Tout d'abord, il peut être beaucoup plus facile de comprendre le comportement de l'estimateur ou de la statistique de test d'intérêt à partir des paramètres d'une surface de réponse plutôt qu'à partir de plusieurs tableaux remplis de chiffres. Ensuite, si la surface de réponse est correctement spécifiée, elle élimine, ou du moins réduit dans de grandes proportions, ce que Hendry (1984) appelle le problème de **spécificité**. Ce terme signifie que chaque expérience individuelle donne des résultats seulement pour un unique DGP supposé, et n'importe quel ensemble d'expériences Monte Carlo donne des résultats seulement pour un ensemble fini de DGP supposés. Pour d'autres valeurs paramétriques ou d'autres valeurs de  $n$ , le lecteur doit interpoler le résultat à partir des résultats des tableaux, ce qui est



souvent difficile. Par contraste, une surface de réponse correctement spécifiée fournit des résultats pour des familles entières de DGP plutôt que pour des valeurs spécifiques choisies par l'expérimentateur. Le revers de la médaille, naturellement, est que la surface de réponse doit être correctement spécifiée, et cela n'est pas toujours une tâche facile.

Une des caractéristiques les plus intéressantes des surfaces de réponse, qui les distingue de la plupart des autres applications des modèles de régression en économie, est que les données sont générées par l'expérimentateur. Ainsi, si les données ne sont pas suffisamment informatives, il y a toujours une solution facile: exécuter davantage d'expériences pour obtenir davantage de données. Dans la plupart des cas, chaque point (chaque donnée) de la surface de réponse correspond à une seule expérience Monte Carlo. La variable dépendante est alors une quantité quelconque estimée par l'expérience, telle la moyenne ou l'erreur quadratique moyenne des estimations d'un certain paramètre ou le niveau estimé d'un test. Comme de telles estimations sont normalement accompagnées des estimations de leurs écarts types, des estimations qui devraient être très précises si les expériences comportent un nombre suffisant de répétitions, le chercheur est dans l'obligation d'utiliser les GLS avec une matrice de covariance pleinement spécifiée. Si chaque expérience avait utilisé un ensemble différent de nombres aléatoires, les observations pour la surface de réponse seraient indépendantes, et cette matrice de covariance serait par conséquent diagonale. Si les mêmes nombres aléatoires étaient utilisés dans plusieurs expériences, peut-être pour augmenter la précision avec laquelle les différences entre les valeurs paramétriques seraient estimées, la matrice de covariance serait naturellement non diagonale, mais la forme de la non-diagonalité serait connue, et l'on pourrait estimer la matrice de covariance assez facilement.

Afin de rendre les remarques précédentes plus concrètes, notons  $\psi$  la quantité d'intérêt. Elle doit être une fonction de la taille de l'échantillon  $n$  et des paramètres qui caractérisent le DGP, que nous pouvons noter sous forme vectorielle  $\alpha_0$ . Nous modéliserons cette fonction par  $\Psi(n, \alpha_0, \gamma)$ , où  $\Psi$  est une forme fonctionnelle spécifique qui dépend d'un vecteur paramétrique  $\gamma$ , qui sera estimé. La surface de réponse que nous essayons d'estimer est alors caractérisée par  $\Psi(n, \alpha_0, \gamma_0)$  pour un vecteur approprié  $\gamma_0$  quelconque. Cette expression nous indique comment  $\psi$  varie suite à des changements de  $n$  et de  $\alpha_0$ . La  $i^{\text{ième}}$  expérience génère une *estimation*  $\hat{\psi}_i$  et un écart type associé  $\hat{\sigma}(\hat{\psi}_i)$ . L'estimation  $\hat{\psi}_i$  peut être soit une simple moyenne sur  $N$  répétitions (comme nous l'avons vu dans la Section 21.5, ceci est le cas même si les variables antithétiques ont été utilisées, sauf qu'il s'agit alors d'une moyenne sur  $N$  doubles répétitions), soit une estimation CV, provenant probablement soit de la régression (21.14) soit de la régression (21.15). Quoi qu'il en soit, si le nombre de régressions par expérience est raisonnablement grand, nous pouvons être assurés que  $\hat{\psi}_i$  est pratiquement normal avec une espérance  $\Psi(n, \alpha_0, \gamma_0)$  et un écart type  $\sigma(\hat{\psi}_i)$ , et ce dernier sera bien estimé par  $\hat{\sigma}(\hat{\psi}_i)$ . Ainsi la

régression de la surface de réponse est

$$\hat{\psi}_i = \Psi(n, \alpha_0, \gamma) + v_i, \quad v_i \sim N(0, \hat{\sigma}^2(\hat{\psi}_i)), \quad i = 1, \dots, M, \quad (21.21)$$

où  $M$  est le nombre d'expériences et par conséquent le nombre d'observations pour la surface de réponse. En transformant (21.21) pour éliminer l'hétéroscédasticité, nous obtenons

$$\frac{\hat{\psi}_i}{\hat{\sigma}(\hat{\psi}_i)} = \frac{\Psi(n, \alpha_0, \gamma)}{\hat{\sigma}(\hat{\psi}_i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \quad i = 1, \dots, M. \quad (21.22)$$

Les arguments précédents suggèrent que, à condition que le nombre de répétitions par expérience soit raisonnablement grand, la spécification des aléas dans (21.22) avec la  $N(0, 1)$  devrait être une approximation extrêmement bonne. Cependant, certains auteurs ont avancé l'idée que le nombre de répétitions par expérience peut être beaucoup plus faible quand il s'agit d'estimer une surface de réponse que lorsqu'il s'agit d'exécuter des expériences Monte Carlo plus conventionnelles. Par exemple, Engle, Hendry, et Trumble (1985) utilisent seulement 21 répétitions par expérience. Il est vrai que l'on peut souvent estimer les paramètres  $\gamma$  de  $\Psi(n, \alpha_0, \gamma)$  avec une grande précision même quand  $N$  est petit, à condition que  $M$  soit suffisamment grand, parce qu'un grand nombre d'expériences peut compenser des résultats imprécis provenant de chaque expérience individuelle. Cependant, deux problèmes peuvent survenir quand  $N$  est petit. Tout d'abord, la distribution de  $\hat{\psi}_i - \Psi(n, \alpha_0, \gamma)$  peut différer assez significativement de la distribution normale, et  $\hat{\sigma}(\hat{\psi}_i)$  peut être une piètre estimation de  $\sigma(\hat{\psi}_i)$ . Ceci signifie que l'inférence basée sur (21.22) peut être problématique. En second lieu, si  $\hat{\psi}_i$  n'est pas une estimation précise, il peut être difficile de spécifier la forme fonctionnelle de  $\Psi(n, \alpha_0, \gamma)$ . Comme nous le verrons par la suite, le plus gros problème en pratique lié à l'utilisation des surfaces de réponse est que la forme de  $\Psi(n, \alpha_0, \gamma)$  n'est généralement pas connue a priori. La présence d'estimations précises  $\hat{\psi}_i$  peut être d'un grand secours dans la spécification de la forme fonctionnelle de  $\Psi(n, \alpha_0, \gamma)$ .

La meilleure manière d'expliquer l'estimation des surfaces de réponse est de fournir un exemple concret. Le problème que nous étudierons a l'aspect de celui traité dans la section précédente et a été aussi utilisé comme exemple par Hendry (1984), à savoir le biais de l'estimation OLS  $\hat{\beta}$  dans le modèle autorégressif stationnaire (21.16). Il s'agit naturellement d'un problème qui a été largement étudié par d'autres méthodes pendant longtemps; consulter, par exemple, Hurwicz (1950). Il est en réalité trop simple pour être l'objet d'une expérience Monte Carlo, parce qu'on peut calculer le biais de  $\hat{\beta}$  analytiquement, comme dans Sawa (1978), à condition que les aléas soient normaux, comme nous le supposons. Cependant, les calculs demandés ne sont en au-

cune manière triviale, et il n'existe aucune formule rapidement interprétable qui relie le biais de  $\hat{\beta}$  aux valeurs de  $\beta_0$  et  $n$ .<sup>4</sup>

Phillips (1977) essaie de dériver une telle formule à partir de la méthode des développements asymptotiques. Ici nous essayons de procéder de la sorte en estimant une surface de réponse, en utilisant des résultats à partir des expériences Monte Carlo pour obtenir des points (données).

Nous avons tout d'abord généré des données à partir de 390 expériences, en faisant varier  $\beta_0$  de  $-0.95$  à  $0.95$  par incrément de  $0.05$  et, pour chaque  $\beta_0$ , en essayant  $n = 16, 25, 36, 49, 64, 81, 100, 150, 200$ , et  $400$ . Nous n'avons pas utilisé volontairement des valeurs de  $|\beta_0|$  supérieures à  $0.95$  parce qu'il serait sûrement difficile de caractériser le comportement de  $\hat{\beta}$  par une surface de réponse unique aussi bien pour le cas stationnaire que le cas de racine unitaire, et nous avons vu que des phénomènes étranges commencent à survenir quand  $|\beta_0| \rightarrow 1$  (rappelons-nous de la Figure 21.3). Le nombre de répétitions utilisé dans les expériences était relativement faible: 2000 pour  $n = 16$  et  $25$ ; 1000 pour  $n = 36$  et  $49$ ; 500 pour  $n = 64, 81$  et  $100$ ; et 250 pour  $n = 150, 200$ , et  $400$ . Nous avons utilisé plus de répétitions pour des valeurs inférieures de  $n$  parce que les estimations CV de la moyenne de  $\hat{\beta}$  étaient beaucoup moins précises pour un nombre donné de répétitions. La régressande pour la surface de réponse était l'estimation CV de la moyenne de  $\hat{\beta}$ , moins  $\beta_0$ , divisée par l'écart type estimé de la moyenne de  $\hat{\beta}$ , le tout obtenu à partir de la régression (21.14). Notons que les estimations de la moyenne de  $\hat{\beta}$  étaient très précises: les écarts types estimés variaient de  $.000190$  (pour  $\beta_0 = .05$  et  $n = 400$ ) à  $.002813$  (pour  $\beta_0 = .90$  et  $n = 16$ ).

Il fut facile de générer des données, mais la spécification de la surface de réponse fut beaucoup plus délicate. Dans ce cas, nous pouvons écrire l'équation (21.22) comme

$$\frac{\hat{\beta}_i - \beta_0}{\hat{\sigma}(\hat{\beta}_i)} = \frac{\Psi(n, \beta_0, \gamma)}{\hat{\sigma}(\hat{\beta}_i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \quad i = 1, \dots, 390,$$

où  $\Psi(n, \beta_0, \gamma)$  est la fonction biais que nous essayons d'estimer. La théorie asymptotique nous enseigne que  $\Psi(n, \beta_0, \gamma)$  tend vers zéro quand  $n \rightarrow \infty$ . Ceci signifie qu'il ne devrait y avoir aucun terme constant et que tous les régresseurs devraient être divisés par une certaine puissance positive de  $n$ . Malgré tout, ceci laisse encore une grande plage de possibilités. Nous avons

<sup>4</sup> Notons que des problèmes étroitement liés, tels que les propriétés des  $t$  de Student pour ce modèle, ne peuvent pas être traités analytiquement. Nankervis et Savin (1988) utilisent une gamme extrêmement complète d'expériences Monte Carlo pour étudier les propriétés des  $t$  de Student dans une version légèrement plus compliquée de (21.16) dans laquelle il faut estimer un terme constant. Cet article est l'un des meilleurs exemples disponibles des méthodes Monte Carlo en application.

tout d'abord estimé des fonctions de biais très simples<sup>5</sup>

$$\begin{aligned}\Psi(n, \beta_0, \gamma) &= -1.6890 \, n^{-1} \beta_0 \\ &\quad (0.0108) \\ s &= 1.8038, \quad DW = 1.0322, \quad \bar{R}^2 = 0.9844.\end{aligned}\tag{21.23}$$

Hendry (1984) a estimé une fonction de cette forme en tant que première approximation mais l'a trouvée très insatisfaisante. Ces résultats sont également très peu satisfaisants. Bien que le  $\bar{R}^2$  soit très élevé, ce qui implique que  $n^{-1}\beta_0$  explique un très grand pourcentage de la variation totale de  $\hat{\beta} - \beta_0$ , l'écart type estimé de l'équation est bien supérieur à sa valeur théorique de 1, et la statistique Durbin-Watson est nettement inférieure à 2. Puisque les données étaient classées par  $n$  (toutes les observations pour les  $n = 16$  premières, puis toutes les observations pour  $n = 25$ , et ainsi de suite), la faible valeur de la statistique DW suggère fortement que la relation entre le biais et la taille d'échantillon est mal spécifiée.

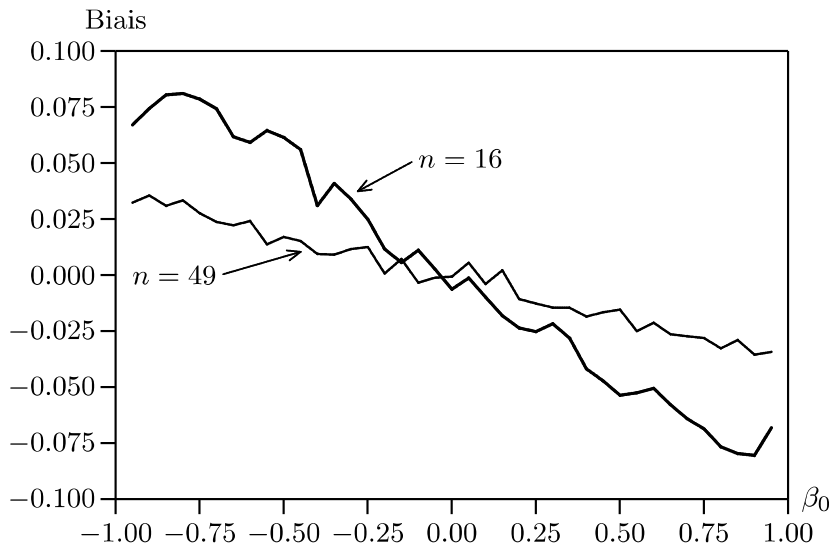
La prochaine étape évidente était d'ajouter à (21.23) les termes associés à des puissances de  $\beta_0$  divisés par les puissances de  $n$ . La littérature sur les développements asymptotiques, par exemple Phillips (1977), suggère que l'on devrait utiliser des puissances multiples d'un demi. Ainsi, on pourrait essayer d'estimer un modèle général de la forme

$$\Psi(n, \beta_0, \gamma) = \sum_{a=1}^6 \sum_{b=1}^6 \gamma_{ab} n^{-a/2} \beta_0^{b/2} \tag{21.24}$$

et ensuite essayer de le simplifier et annulant de nombreux  $\gamma_{ab}$ . On voudrait laisser  $a$  et  $b$  s'incrémenter jusqu'à 6 parce que Hendry (1984) a semblé mettre en évidence le fait que  $\beta_0^3/n^3$  appartenait à  $\Psi(n, \beta_0, \gamma)$ . Ce modèle doit forcément mieux s'ajuster que (21.23), mais les estimations seront extrêmement imprécises parce qu'il y a 36 régresseurs potentiels de la forme  $n^{-a/2} \beta_0^{b/2}$ , et certains d'entre eux seront fortement colinéaires. Par conséquent nous avons considéré que la spécification d'une surface de réponse de cette manière était impossible. Il n'y avait tout simplement aucun moyen pertinent d'obtenir un modèle plus économe à partir du modèle général (21.24). Si cette approche est insatisfaisante dans ce cas très simple, où le DGP ne comprend qu'un seul paramètre, elle sera totalement insatisfaisante en général.

Par conséquent, nous avons choisi une approche radicalement différente, en utilisant des méthodes graphiques pour voir à quoi  $\Psi(n, \beta_0, \gamma)$  doit ressembler. Cette approche fut utilisée avec succès. Elle ne fut possible que parce

<sup>5</sup> Ces expériences étaient à exécutées pour la première fois en 1988 et nécessitaient environ 16 heures sur un ordinateur de type 286. Puisqu'ils auraient pris moins de dix minutes sur un PC 486, il aurait été possible d'utiliser davantage de répétitions.



**Figure 21.4** Estimations Monte Carlo du biais

que nos estimations de  $\hat{\beta} - \beta_0$  étaient très précises, ce qui garantissait une lisibilité immédiate des graphiques illustrant les variations de  $\hat{\beta} - \beta_0$  en fonction de  $\beta_0$  pour des valeurs diverses de  $n$ , et celles de  $\hat{\beta} - \beta_0$  en fonction de  $n$  pour des valeurs diverses de  $\beta_0$ . C'est une raison pour ne pas utiliser de petites valeurs de  $N$  dans des expériences Monte Carlo destinées à l'estimation des surfaces de réponse.

La Figure 21.4 illustre les graphes de  $\hat{\beta} - \beta_0$  en fonction de  $\beta_0$  pour  $n = 16$  et  $n = 49$ . Il est évident que la relation est fondamentalement linéaire et symétrique autour de zéro, sauf que pour  $n = 16$  (et évidemment pour d'autres valeurs plus petites de  $n$ ) il y a une inversion assez brutale de la pente pour de grandes valeurs absolues de  $\beta_0$ . Il est aussi évident à partir de la figure que la relation entre  $\hat{\beta} - \beta_0$  et  $\beta_0$  devient moins prononcée quand  $n$  augmente; la relation pour  $n = 400$  (non présentée pour éviter de saturer la figure) était presque plate.

Le comportement évident dans la Figure 21.4 de la relation entre  $\hat{\beta} - \beta_0$  et  $\beta_0$  pour de grandes valeurs absolues de  $\beta_0$  suggère que l'on pourrait vouloir ajouter des fonctions de  $\beta_0^3$  dans  $\Psi(n, \beta_0, \gamma)$ . Cependant, il existe d'autres fonctions de  $\beta_0$  qui pourraient tout aussi bien traduire la pente évidente dans la figure, notamment  $\beta_0/(1 - \beta_0^2)$  et  $\beta_0/(1 - \beta_0^2)^{1/2}$ . En régressant  $\hat{\beta} - \beta_0$  sur  $\beta_0$  et sur un autre régresseur parmi  $\beta_0^3$ ,  $\beta_0/(1 - \beta_0^2)$ , et  $\beta_0/(1 - \beta_0^2)^{1/2}$  pour des valeurs diverses de  $n$ , nous avons conclu que  $\beta_0/(1 - \beta_0^2)^{1/2}$  expliquait le mieux la relation observée entre  $\hat{\beta} - \beta_0$  et  $\beta_0$ .

Des graphes similaires et des régressions préliminaires ont suggéré que  $n^{-1}$  et  $n^{-3/2}$  expliquaient ensemble pratiquement toute la relation entre  $\hat{\beta} - \beta_0$  et la taille de l'échantillon, mais qu'au contraire  $n^{-1/2}$  et  $n^{-2}$  ne jouaient

aucun rôle. Ainsi, nous avons choisi à titre d'essai la spécification

$$\begin{aligned}\Psi(n, \beta_0, \gamma) = & n^{-1}(\gamma_1 + \gamma_2\beta_0 + \gamma_3\beta_0/(1 - \beta_0^2)^{1/2}) \\ & + n^{-3/2}(\gamma_4 + \gamma_5\beta_0 + \gamma_6\beta_0/(1 - \beta_0^2)^{1/2}).\end{aligned}\quad (21.25)$$

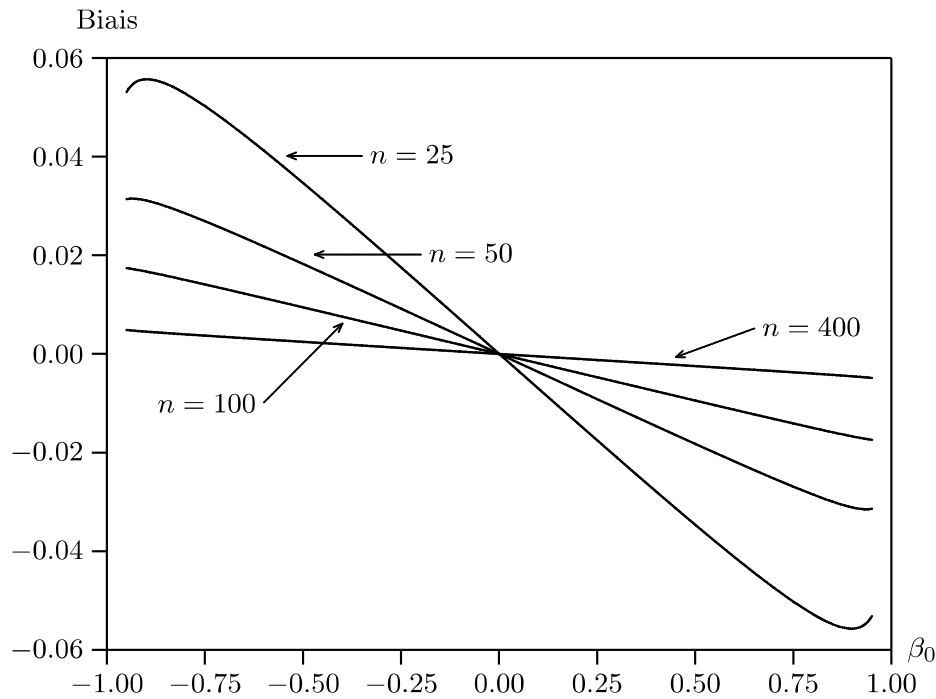
Celle-ci est excessivement plus simple que (21.24). Quand (21.25) fut estimée, nous avons trouvé que  $\tilde{\gamma}_1$ ,  $\tilde{\gamma}_4$ , et  $\tilde{\gamma}_5$  étaient conjointement non significatifs, bien que  $\tilde{\gamma}_4$  était individuellement significatif à un niveau de 5%. Puisqu'il est difficile de voir pourquoi  $\hat{\beta}$  devrait être biaisé quand  $\beta_0 = 0$ , et puisque par contraste avec  $\tilde{\gamma}_4$  les trois autres paramètres significatifs étaient fortement significatifs, nous avons décidé sur la base de ces résultats de contraindre  $\gamma_1$ ,  $\gamma_4$ , et  $\gamma_5$  dans (21.25) à zéro. Nos estimations du modèle résultant étaient

$$\begin{aligned}\Psi(n, \beta_0, \gamma) = & -1.9223 \frac{n^{-1}\beta_0}{(0.0173)} - 0.1066 \frac{n^{-1}\beta_0}{(0.0149)(1 - \beta_0^2)^{1/2}} \\ & + 1.3509 \frac{n^{-3/2}\beta_0}{(0.0608)(1 - \beta_0^2)^{1/2}}\end{aligned}\quad (21.26)$$

$$s = 1.0628, \quad DW = 1.8649, \quad \bar{R}^2 = 0.9946.$$

Ces résultats apparaissent être très bons. Les trois paramètres sont très significatifs, l'écart type de la régression est légèrement supérieur à 1, mais pas de manière significative au niveau 5%, et la statistique DW n'est pas significativement inférieure à 2. Les tests d'asymétrie et d'aplatissement n'ont pas décelé ces phénomènes. De plus, quand d'autres fonctions diverses de  $\beta_0$  et  $n$ , telles que  $n^{-1}\beta_0/(1 - \beta_0^2)$ ,  $n^{-1}\beta_0^3$ ,  $n^{-3/2}\beta_0/(1 - \beta_0^2)$ ,  $n^{-3/2}\beta_0^3$ ,  $n^{-2}\beta_0$ , et  $n^{-2}\beta_0/(1 - \beta_0^2)^{1/2}$ , étaient intégrées à  $\Psi(n, \beta_0, \gamma)$ , elles étaient individuellement et conjointement non significatives, et les trois régresseurs dans (21.26) sont restés individuellement significatifs. Pour des tailles d'échantillon dans la gamme examinée, les valeurs prédites par (21.26) sont très proches des valeurs exactes tabulées par Sawa (1978), bien que l'équation semble prédire un résultat quelque peu trop biaisé pour de faibles valeurs de  $n$ .

Nous concluons que la surface de réponse (21.26) fournit une bonne approximation, bien que non parfaite, pour la fonction de biais  $\Psi(n, \beta_0, \gamma)$  sur l'intervalle  $n = 16$  à  $n = \infty$  et  $\beta_0 = -0.95$  à  $\beta_0 = 0.95$ . Cependant, cela peut ne pas être le cas pour de très petites valeurs de  $n$  et pour des valeurs de  $|\beta_0|$  supérieures à 0.95. Un ensemble d'expériences beaucoup plus coûteux et selon toute vraisemblance une surface de réponse considérablement plus compliquée seraient nécessaires si nous décidions de traiter de façon adéquate ces cas. Cette surface de réponse est illustrée comme une fonction de  $\beta_0$  pour des valeurs variées de  $n$  dans la Figure 21.5. Les tendances du biais à diminuer fortement quand  $n$  augmente, et à augmenter avec  $|\beta_0|$  sauf pour une légère diminution pour de grandes valeurs de  $|\beta_0|$  sont relativement évidentes sur la figure.



**Figure 21.5** Estimations de biais par surface de réponse

Dans toutes les estimations rapportées jusqu'ici, nous avons utilisé les estimations CV de  $\hat{\beta}$ . Il aurait été aussi possible d'utiliser les estimations naïves de  $\hat{\beta}$ . La surface de réponse estimée quand nous avons procédé de la sorte était

$$\begin{aligned} \Psi(n, \beta_0, \gamma) = & - \frac{1.9272}{(0.0366)} n^{-1} \beta_0 - \frac{0.1306}{(0.0274)} n^{-1} \frac{\beta_0}{(1 - \beta_0^2)^{1/2}} \\ & + \frac{1.4983}{(0.1141)} n^{-3/2} \frac{\beta_0}{(1 - \beta_0^2)^{1/2}} \end{aligned} \quad (21.27)$$

$$s = 1.0811, \quad DW = 1.8606, \quad \bar{R}^2 = 0.9763.$$

Ces résultats sont très similaires à ceux utilisés pour les estimations CV mais sont moins bons à tous les égards. Les écarts types associés aux estimations paramétriques sont généralement environ deux fois plus grands, et indiquent qu'en moyenne, l'usage des variables de contrôle revient approximativement à quadrupler le nombre de répétitions. La valeur légèrement supérieure de  $s$  indique probablement que la surface de réponse s'ajuste légèrement moins bien pour les petites valeurs de  $n$ . L'usage des variables de contrôle améliore davantage les estimations de  $\hat{\beta}$  pour des valeurs importantes de  $n$ . Ainsi, la surface de réponse (21.26), qui utilise les estimations CV, pondère les résultats des expériences avec des valeurs importantes de  $n$  plus lourdement que ne le fait la surface de réponse (21.27) qui utilise des estimations naïves. Ainsi,

nous nous attendons à ce que (21.27) s'ajuste moins bien que (21.26), comme c'est le cas, si la surface de réponse est moins performante pour des tailles d'échantillon plus petites.

Cet exemple concerne l'estimation d'une fonction de biais. L'estimation des fonctions de MSE, ou des fonctions de niveau ou de puissance pour les statistiques de test, est conceptuellement similaire, bien que certains détails soient naturellement différents. Si la variable dépendante est le niveau ou la puissance d'une statistique de test, que nous pouvons noter  $p$ , alors cette variable dépendante doit varier entre 0 et 1, et la transformation logit

$$\Lambda(p) = \log\left(\frac{p}{1-p}\right)$$

peut être utile. La justification de cette transformation est que  $\Lambda(p)$  peut varier entre plus et moins l'infini, ce qui facilite la spécification d'une surface de réponse comme fonction linéaire. Pour l'essentiel, nous estimerions alors un modèle logit sur des données groupées. (Consulter le Chapitre 15).

Nous croyons que l'exemple précédent est très révélateur. Il illustre combien peuvent être utiles les surfaces de réponse grâce à leur capacité à synthétiser une grande quantité de résultats expérimentaux en un ensemble relativement simple d'estimations comme (21.26), que l'on peut alors représenter graphiquement comme dans la Figure 21.5. Il illustre aussi les difficultés pratiques de spécification d'une surface de réponse. L'approche de la surface de réponse ne sera pas opérationnelle si le DGP est caractérisé par plusieurs paramètres qui affectent les quantités étudiées, parce qu'il sera tout simplement trop difficile de spécifier la surface de réponse dans un tel cas, du moins s'il y a une quelconque interaction entre les différents paramètres. Des méthodes graphiques telles que celles employées peuvent être extrêmement bénéfiques pour la spécification d'une surface de réponse, mais elles ont leurs limites, et il semble malheureusement peu probable qu'elles seront efficaces quand le DGP comporte de nombreux paramètres qui interagissent de façon compliquée.

## 21.8 LE BOOTSTRAP ET LES MÉTHODES CONNEXES

Jusqu'à présent, nous avons porté notre attention sur les expériences Monte Carlo "conventionnelles" dans lesquelles le chercheur spécifie pleinement le DGP pour chaque expérience. Bien que de telles expériences puissent être utilisées comme compléments à des parties précises du travail empirique et sont parfois employées à profit de cette manière, elles sont beaucoup plus communément employées pour suppléer le travail théorique sur les propriétés des estimateurs et des statistiques de test. Par contraste, la technique connue sous le nom du **bootstrap** est typiquement conçue pour être utilisée dans le



contexte du travail empirique. Comme le nom le suggère, l'idée du bootstrap<sup>6</sup> est d'utiliser le seul ensemble de données disponible pour créer une sorte d'expérience Monte Carlo dans laquelle les données elles-mêmes sont utilisées pour approximer la distribution des aléas ou d'autres quantités aléatoires du modèle. Le nom est censé exprimer l'idée que les données disponibles devraient fournir suffisamment d'information sur leur distribution. Cette idée est mise en œuvre par l'exécution d'une sorte d'expérience Monte Carlo dans laquelle les aléas ou les autres quantités aléatoires sont habituellement des tirages non pas d'une distribution supposée, telle que la normale, mais plutôt à partir de la distribution empirique de leurs contreparties d'échantillon. L'obtention d'échantillons artificiels de cette manière est un cas particulier de ce que l'on appelle **rééchantillonnage**; consulter Efron (1979).

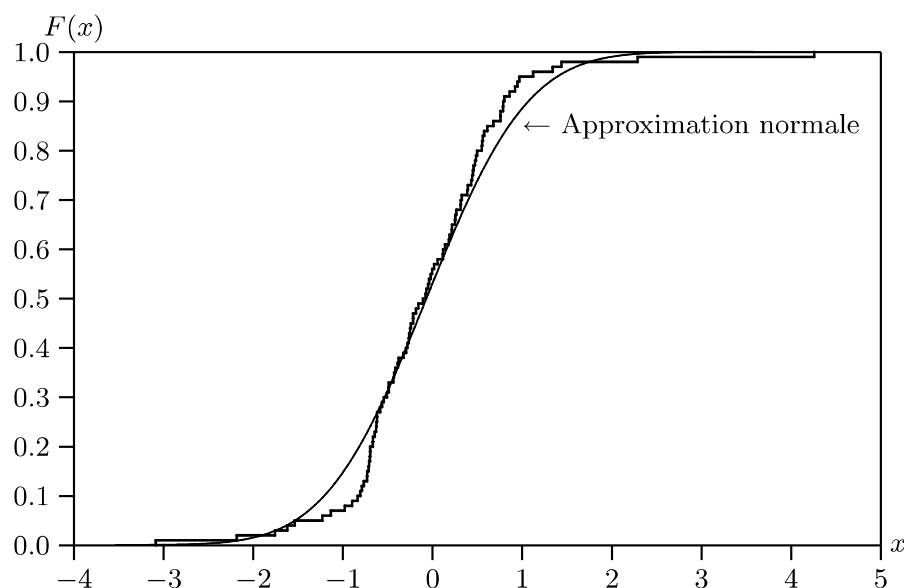
Nous avons rencontré pour la première fois la **fonction de distribution empirique**, ou **EDF**, dans la Section 4.5. Si nous notons  $\{y_t\}_{t=1}^n$  un échantillon de taille  $n$ , où les  $y_t$  sont des réalisations des variables aléatoires indépendantes, alors la EDF est la fonction de répartition

$$\hat{F}^n(x) \equiv \frac{1}{n} \sum_{t=1}^n I_{(-\infty, x)}(y_t),$$

où la fonction indicatrice  $I$  associée à l'intervalle  $(-\infty, x)$  est simplement une fonction qui prend la valeur 1 si son argument appartient à l'intervalle, et 0 sinon. Ainsi, une EDF est une fonction en escalier, la hauteur de chaque marche étant  $1/n$ , et la largeur étant la différence entre deux valeurs successives de  $y_t$  quand ces dernières sont classées par ordre croissant. Si deux ou plusieurs observations sont identiques, événement associé à la probabilité nulle si la densité des  $y_t$  est continue, il peut y avoir des escaliers qui ont une hauteur multiple entier de  $1/n$ . La EDF pour un ensemble particulier de 100 observations sur une variable aléatoire  $y$  est illustrée dans la Figure 21.6; à titre de comparaison, une distribution normale avec les mêmes espérance et variance est aussi reportée.

Supposons que l'on ait calculé des statistiques  $\theta(\mathbf{y})$  quelconques à partir d'un ensemble de données  $y_t$ ,  $t = 1, \dots, n$ , noté sous forme vectorielle  $\mathbf{y}$ ; dans la pratique, on pourrait calculer de nombreuses statistiques différentes, mais pour des raisons de simplicité, nous ne traiterons seulement que l'une d'entre elles. Si la distribution en échantillon fini de  $\theta(\mathbf{y})$  est connue, ou si une bonne approximation asymptotique est disponible, le recours au bootstrap est inutile. Si, cependant, ce n'est pas le cas, une manière d'approximer la distribution de  $\theta(\mathbf{y})$  est d'appliquer le bootstrap à cet ensemble de données. Pour cela, on doit tirer un certain nombre d'**échantillons bootstrap**, disons  $B$ , chacun de taille  $n$ , à partir de la distribution des données observées. Ce rééchantillonnage

<sup>6</sup> Un "bootstrap" en anglais est un tirant de botte. L'expression "to pull oneself up by one's bootstraps" signifie "se faire tout seul".



**Figure 21.6** Fonction de répartition empirique basée sur 100 observations

est réalisé avec *remise*. Ainsi, chaque échantillon bootstrap contiendra certaines des  $n$  observations d'origine plus d'une fois, et d'autres pas du tout, et ce de manière tout à fait aléatoire. Le tirage d'un échantillon bootstrap est très facile. Notons  $y_j^*(i)$  la  $j^{\text{ième}}$  observation du  $i^{\text{ième}}$  échantillon bootstrap, où  $i = 1, \dots, B$ . Pour obtenir  $y_j^*(i)$ , nous générons tout d'abord un nombre pseudo-aléatoire à partir de la distribution  $U(0, 1)$ , l'utilisons pour générer un entier aléatoire  $k$  qui prend les valeurs  $1, \dots, n$  avec équiprobabilité, et ensuite initialisons  $y_j^*(i)$  à  $y_k$ . En répétant cette opération  $n$  fois, nous obtenons un échantillon bootstrap complet, disons  $\mathbf{y}^*(i)$ . Nous calculons ensuite  $\theta(\mathbf{y}^*(i))$  et sauvegardons le résultat. L'opération entière est alors répétée pour  $i = 1, \dots, B$  échantillons bootstrap, à la fin de laquelle nous obtenons  $B$  statistiques  $\theta(\mathbf{y}^*(i))$ . Ces statistiques sont à leur tour utilisées pour estimer n'importe quelle caractéristique de la distribution de  $\theta(\mathbf{y})$  à laquelle on pourrait s'intéresser.

Le paragraphe précédent a esquissé l'idée de base du bootstrap, que l'on doit à Efron (1979). Des références relativement accessibles sont Efron (1982), Efron et Gong (1983), et Efron et Tibshirani (1986). Des références plus théoriques sont Bickel et Freedman (1981), Freedman (1981), et Hall (1987). La littérature est devenue très importante et parfois très technique au cours des dernières années, et nous n'effectuerons aucune tentative ici pour l'examiner.

Illustrons maintenant l'usage du bootstrap dans un cas simple. Considérons les données illustrées dans la Figure 21.6. On peut facilement voir à partir de la figure que ces données sont des tirages d'une distribution comportant des queues plus grosses que la normale. Une distribution normale

avec les mêmes espérance et variance que les données est illustrée dans la figure, et il est évident que les valeurs les plus importantes dans chaque queue de l'échantillon auraient dû survenir avec une probabilité extrêmement faible avec la distribution normale. Un chercheur pourrait par conséquent s'inquiéter et se demander si les inférences basées sur des estimations et les intervalles de confiance issus du cas normal seraient valables dans ce cas. Une manière de voir si de telles inquiétudes sont fondées est d'appliquer le bootstrap aux statistiques d'intérêt.

Considérons l'espérance des  $y_t$ . La moyenne d'échantillon est  $-0.0701$ , avec un écart type de  $0.0889$ . Ainsi, l'intervalle de confiance habituel à 95% basé sur la distribution du  $t$  de Student à 99 degrés de liberté est  $(-0.2464, 0.1062)$ . Nous avons calculé 10,000 échantillons bootstrap comme ceux décrits précédemment, et ainsi obtenu 10,000 moyennes estimées,  $\mu^*(i)$ . Ce choix de  $B$  dépasse celui nécessaire dans la plupart des cas, et garantit une erreur expérimentale très faible. Il y a plusieurs manières d'obtenir des intervalles de confiance bootstrap à partir de la distribution des  $\mu^*(i)$ ; consulter Efron et Tibshirani (1986) pour une introduction et Tibshirani (1988) pour des méthodes plus avancées. La première étape consiste à trier les moyennes bootstrap  $\mu^*(i)$  par ordre croissant,  $\mu^*(1)$  étant la plus faible et  $\mu^*(B)$  la plus forte. Si la distribution des  $\mu^*(i)$  est approximativement symétrique, on peut alors utiliser ce qui est appelé **méthode des centiles**. Supposons que nous voulions un intervalle de confiance à 95%. Alors nous choisissons simplement

$$\frac{1}{2}(\mu^*(250) + \mu^*(251))$$

comme limite inférieure de notre intervalle de confiance et

$$\frac{1}{2}(\mu^*(9750) + \mu^*(9751))$$

comme limite supérieure. Ces valeurs sont choisies de sorte qu'exactement 2.5% des répétitions bootstrap produisent des  $\mu^*(i)$  inférieures à la limite inférieure et 2.5% produisent des  $\mu^*(i)$  supérieures à la limite supérieure de l'intervalle de confiance. L'utilisation de la méthode des centiles pour les données de la Figure 21.6 fournit un intervalle de confiance pour la moyenne des  $y_t$  égal à  $(-0.2387, 0.1053)$ , très similaire à l'intervalle basé sur la distribution du  $t$  de Student.

Si la distribution des  $\mu^*(i)$  n'est pas symétrique, on peut ne pas vouloir utiliser la méthode des centiles, parce qu'elle n'est plus optimale pour omettre le même nombre de  $\mu^*(i)$  à partir de chaque queue de leur EDF si nous voulons que l'intervalle de confiance soit aussi court que possible. Une approche simple consiste à minimiser la quantité

$$\frac{1}{2}(\mu^*(l + .95B) + \mu^*(l + .95B + 1)) - \frac{1}{2}(\mu^*(l - 1) + \mu^*(l))$$

par rapport à l'entier positif  $l < .05B$ .<sup>7</sup> Ainsi, l'objectif consiste à trouver l'intervalle le plus court possible comprenant 95% des  $\mu^*(i)$ . Quand la EDF des  $\mu^*(i)$  est asymétrique, cette **méthode des centiles modifiée** tendra à déplacer l'intervalle de confiance loin de la queue la plus longue de la distribution, parce qu'en éliminant des observations d'un côté et en les additionnant de l'autre côté, cela réduira la longueur de l'intervalle de confiance estimé. Pour les données de la Figure 21.6, la méthode des centiles modifiée fournit des résultats très similaires à ceux de la méthode des centiles ordinaire et à la méthode basée sur la théorie normale usuelle: l'intervalle de confiance à 95% est  $(-0.2399, 0.1031)$ .

Puis, dans cet exemple, le bootstrap a principalement servi à nous rassurer que les méthodes conventionnelles d'inférence concernant la moyenne des  $y_t$  sont vraisemblablement très fiables pour cet ensemble de données, en dépit de l'apparent excès de kurtosis relatif au cas normal. Mais la même procédure pourrait être employée pour étudier la distribution de *n'importe quelle* statistique  $\theta(\mathbf{y})$  à laquelle nous nous intéresserions, et parmi elles celles pour lesquelles les méthodes les plus conventionnelles d'inférence sont difficiles ou impossibles. C'est dans de tels cas que le bootstrap peut être particulièrement utile.

La méthode du bootstrap qui vient juste d'être décrite peut évidemment être modifiée de différentes façons. On pourrait, par exemple, lisser quelque peu la EDF des  $y_t$  et tirer des échantillons bootstrap à partir de la EDF lissée à la place de la EDF ordinaire. Si l'on connaissait ou était prêt à supposer la forme de la distribution des  $y_t$ , on pourrait utiliser ce qui est souvent appelé **bootstrap paramétrique**, dans lequel les données sont utilisées pour estimer la densité des  $y_t$ , et les échantillons bootstrap sont alors générés à partir de cette densité estimée. Le bootstrap paramétrique ressemble ainsi à une expérience Monte Carlo ordinaire dans laquelle les paramètres du DGP sont estimés à partir de l'ensemble de données d'intérêt.

Il existe des caractéristiques particulières des méthodes bootstrap appliquées aux modèles de régression. Supposons que le modèle soit

$$y_t = x_t(\boldsymbol{\beta}) + u_t, \quad t = 1, \dots, n, \quad (21.28)$$

où toutes les variables dont  $x_t(\boldsymbol{\beta})$  dépend sont supposées fixes ou du moins indépendantes de tous les  $u_t$ . Si ces derniers sont supposés i.i.d., l'approche naturelle est d'appliquer le bootstrap aux résidus. Avec cette approche, on estime tout d'abord le modèle (21.28) par NLS, afin d'obtenir des estimations paramétriques  $\hat{\boldsymbol{\beta}}$  et des résidus,  $\hat{u}_1$  jusqu'à  $\hat{u}_n$ , et on génère ensuite des échantillons bootstrap à partir du processus générateur de données

$$y_j(i) = x_j(\hat{\boldsymbol{\beta}}) + u_j^*(i), \quad j = 1, \dots, n, \quad (21.29)$$

<sup>7</sup> Ceci suppose que  $.95B$  est un entier, ce qui sera le cas si  $B$  est un multiple entier de 100.

où les  $u_j^*(i)$  sont des échantillons aléatoires avec remise à partir de  $\hat{u}_1, \dots, \hat{u}_n$ . Si  $x_t(\beta)$  dépend des valeurs passées de  $y_t$ , cette approche reste valable, mais dans (21.29)  $y_1(i), \dots, y_{j-1}(i)$  doit être utilisé à la place des vrais  $y_t$  retardés en calculant  $x_j(\beta)$ . Puisque le modèle (21.28) est non linéaire, le bootstrap peut être assez coûteux, et la technique est par conséquent utilisée tout d'abord avec les modèles linéaires.

Cette approche comporte deux autres problèmes. Le premier est que, comme d'habitude, les résidus  $\hat{u}_t$  tendent à sous-estimer les aléas  $u_t$ . Ceci peut être traité en utilisant les résidus modifiés

$$\tilde{u}_t = \frac{\hat{u}_t}{(1 - \hat{h}_t)^{1/2}} - \frac{1}{n} \sum_{s=1}^n \frac{\hat{u}_s}{(1 - \hat{h}_s)^{1/2}}, \quad (21.30)$$

où

$$\hat{h}_t \equiv \hat{\mathbf{X}}_t (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}_t^\top$$

et  $\hat{\mathbf{X}}$ , comme d'habitude, est la matrice des dérivées de  $x_t(\beta)$  par rapport aux éléments de  $\beta$ , évaluée en  $\hat{\beta}$ . La raison pour laquelle nous voudrions diviser  $\hat{u}_t$  par  $(1 - \hat{h}_t)^{1/2}$  est évidente. Comme nous l'avons vu pour la première fois dans la Section 3.2, dans le cas d'un modèle de régression linéaire à aléas i.i.d.,

$$E(u_t^2) = (1 - h_t) \sigma^2.$$

Par conséquent, la division  $\hat{u}_t$  par  $(1 - h_t)^{1/2}$  fournirait des résidus modifiés ayant précisément la bonne variance. La division par  $(1 - \hat{h}_t)^{1/2}$  est l'analogue naturel de cette procédure pour le cas non linéaire et se justifie par le résultat théorique (5.57) de la Section 5.6. Dans (21.30), nous soustrayons ensuite la moyenne des  $\hat{u}_t / (1 - \hat{h}_t)^{1/2}$ , qui ne sera pas nulle en général, afin de garantir une moyenne nulle aux  $\tilde{u}_t$ ; consulter Weber (1984).

Le second problème avec cette approche du bootstrap est que les aléas  $u_t$  sont supposés indépendamment et identiquement distribués. Quand cette hypothèse est douteuse, une seconde approche peut être utilisée. Dans cette seconde approche, nous rééchantillonons à partir de  $(y_t, x_t(\hat{\beta}))$  plutôt qu'à partir de  $\hat{u}_t$  ou de  $\tilde{u}_t$ . Un élément type de l'échantillon bootstrap est  $(y_k, x_k(\hat{\beta}))$ , où  $k$  est un tirage aléatoire à partir de  $1, \dots, n$ . Dans le cas linéaire, chaque élément de l'échantillon bootstrap est  $(y_k, \mathbf{X}_k)$ , où  $\mathbf{X}_k$  est la  $k^{\text{ième}}$  ligne de la matrice des observations des variables indépendantes. Cette seconde approche est clairement irréalisable si  $x_t(\beta)$  dépend des valeurs retardées de  $y_t$ , puisqu'il est sans pertinence d'utiliser de véritables  $y_t$  retardés, et nous n'avons aucune manière de générer des  $y_t$  retardés à partir du bootstrap. Cependant, elle a l'avantage d'être valable même en présence d'hétéroscédasticité. En effet, cette forme du bootstrap produit des résultats souvent très similaires à ceux provenant de l'usage d'un estimateur de la matrice de covariance robuste à l'hétéroscédasticité.

Aucune de ces approches du bootstrap ne nous permet de traiter des modèles dont les aléas sont supposés autocorrélés mais dont la forme d'autocorrélation est inconnue. Le rééchantillonnage détruit toute sorte de dépendance qu'il peut y avoir dans les données d'origine, de sorte que les résultats du bootstrap peuvent ne pas être très fiables si une telle corrélation constitue un problème.

Les applications des méthodes bootstrap pour les économètres comprennent Freedman et Peters (1984), Korajczyk (1985), Bernard et Veall (1987), et Veall (1987). Les deux premiers articles utilisent le bootstrap pour améliorer les inférences sur des modèles estimés pour lesquels la théorie asymptotique disponible pourrait se révéler peu fiable. Les deux suivants l'utilisent pour estimer les intervalles de confiance pour des prévisions, un sujet souvent extrêmement difficile à réaliser de façon analytique quand la technique de prévision est compliquée. Fair (1980) s'est aussi intéressé à la précision des prévisions et, bien que cet article n'utilise pas le terme, il peut être considéré comme un exemple de bootstrap paramétrique. Raj et Taylor (1989) examinent les propriétés en échantillon fini des statistiques de test basées sur le bootstrap, et Veall (1992) montre comment utiliser le bootstrap pour la sélection de modèle.

Comme les coûts de calcul informatique diminuent, il est vraisemblable que des utilisateurs toujours plus nombreux se tourneront vers des variantes du bootstrap pour traiter des modèles où la théorie asymptotique peut être inadaptée. Ceci soulève la question de la pertinence du bootstrap pour traiter de tels modèles. Excepté peut-être dans certains cas particuliers, la seule manière de répondre à cette question serait d'exécuter des expériences Monte Carlo dont les objets seraient des estimations bootstrap. Malheureusement, cela sera souvent très coûteux, puisque s'il y a  $N$  simulations par expérience et que  $B$  échantillons bootstrap sont nécessaires pour obtenir chaque estimation bootstrap, une seule expérience comporterait un total de  $BN$  estimations. A moins que chaque estimation ne puisse être réalisée très rapidement, une telle expérience pourrait consommer un temps de calcul extrêmement important. Cependant, compte tenu de l'évolution des performances des ordinateurs, nous pouvons certainement nous attendre à voir des études Monte Carlo sur le bootstrap dans des situations qui intéressent les économètres, aussi bien qu'une utilisation plus large du bootstrap dans les travaux appliqués.

## 21.9 CONCLUSION

La publication de cet ouvrage correspond avec la commercialisation d'ordinateurs encore plus puissants que les grosses unités de calcul construites au début des années 80 et dont le prix de vente est tellement faible que tous les bureaux des économètres en seront équipés. Dans ce contexte, les méthodes Monte Carlo devraient selon toute vraisemblance être beaucoup plus utilisées que cela n'a été le cas jusqu'à présent. Des lecteurs et des éditeurs refuseront

d'accepter des résultats basés sur des méthodes destinées à l'estimation et à l'inférence qui ont des propriétés statistiques seulement connues asymptotiquement, quand ils sauront que de meilleures approximations peuvent presque toujours être obtenues compte tenu d'un certain coût de calcul. Certaines formes du bootstrap, qui dans sa version paramétrique ressemble fort aux expériences Monte Carlo les plus conventionnelles sur lesquelles nous nous sommes principalement concentrés, seront ainsi vraisemblablement utilisées de façon automatique comme partie intégrante de nombreux articles empiriques.

## TERMES ET CONCEPTS

bootstrap	méthode des centiles modifiée
bootstrap paramétrique	méthodes Monte Carlo
développements asymptotiques (approximations en échantillon fini)	module (pour générateur congruentiel)
échantillon bootstrap	multiplicateur (pour générateur congruentiel)
expérience Monte Carlo	nombres pseudo-aléatoires
fonction de distribution empirique (EDF)	rééchantillonnage
générateur congruentiel (des nombres pseudo-aléatoires)	répétitions
générateur congruentiel multiplicatif	spécificité (problème de)
générateur de nombres aléatoires	surface de réponse
incrément (pour générateur congruentiel)	techniques de réduction de variance
méthode Box-Muller	valeur d'origine (pour générateur de nombres aléatoires)
méthode de rejet	variables antithétiques
méthode de transformation	variables de contrôle
méthode des centiles	variables pseudo-aléatoires

# Annexe A

## Algèbre Matricielle

### A.1 INTRODUCTION

Comme tous ceux qui ont étudié l'économétrie ou une quelconque autre discipline mathématique le savent, la différence entre un résultat qui semble obscur et difficile, et un résultat qui semble clair et intuitif, provient souvent simplement de la notation utilisée. Dans presque tous les cas, la notation la plus claire rend possible l'utilisation des vecteurs et des matrices. Les lecteurs de ce livre devraient être assez familiers avec l'algèbre matricielle. Cette annexe est destinée à aider ceux qui espèrent se rafraîchir la mémoire et réunir les résultats avec une plus grande facilité. Les lecteurs devraient noter que le Chapitre 1 contient aussi un nombre utile de résultats sur les matrices, en particulier ceux concernant les matrices de projection. Dans cette annexe, des preuves seront données seulement si elles sont courtes ou si elles sont intéressantes. Ceux qui sont intéressés par un traitement plus complet et plus rigoureux peuvent se reporter à Lang (1987).

### A.2 FAITS ÉLÉMENTAIRES CONCERNANT LES MATRICES

Une **matrice**  $\mathbf{A}$  de dimension  $n \times m$  est un tableau rectangulaire de chiffres qui se compose de  $nm$  éléments arrangés dans  $n$  lignes et  $m$  colonnes. Le nom de la matrice est de façon conventionnelle retranscrit en caractères gras. Un élément type de la matrice  $\mathbf{A}$  pourrait être noté  $A_{ij}$  ou  $a_{ij}$ , où  $i = 1, \dots, n$  et  $j = 1, \dots, m$ . Le premier indice désigne toujours la ligne et le second la colonne. Il est parfois nécessaire de montrer explicitement les éléments d'une matrice, dans ce cas ils sont disposés en lignes et en colonnes et entourés par de grands crochets, comme dans

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 5 & 5 \end{bmatrix}.$$

Ici  $\mathbf{B}$  est une matrice de dimension  $2 \times 3$ .

Si une matrice n'a qu'une seule colonne ou une seule ligne, elle est appelée **vecteur**. Il existe deux types de vecteurs, des **vecteurs colonnes** et des **vecteurs lignes**, dont les noms sont explicites. Puisque le premier type est



plus courant que le second, un vecteur qui n'est pas spécifié pour être vecteur ligne sera traité comme un vecteur colonne. Si un vecteur colonne comporte  $n$  éléments, il s'agira d'un vecteur à  $n$  dimensions. Le caractère gras est utilisé pour désigner des vecteurs aussi bien que des matrices. Il est conventionnel d'utiliser des majuscules pour les matrices et des minuscules pour les vecteurs. Cependant, il est parfois nécessaire d'ignorer cette convention.

Si une matrice a le même nombre de colonnes que de lignes, elle est **carrée**. Une matrice carrée  $\mathbf{A}$  est **symétrique** si  $A_{ij} = A_{ji}$  pour tout  $i$  et  $j$ . Des matrices symétriques surviennent très fréquemment en économétrie. Une matrice carrée est **diagonale** si  $A_{ij} = 0$  pour tout  $i \neq j$ ; dans ce cas, les seuls éléments non nuls sont ceux qui forment la **diagonale principale**. Parfois une matrice carrée est composée de zéros au-dessus ou au-dessous de la diagonale principale. Une telle matrice est dite **triangulaire**. Si les éléments non nuls sont tous au-dessus de la diagonale, elle est dite **triangulaire-supérieure**; si les éléments non nuls sont tous au-dessous de la diagonale, elle est dite **triangulaire-inférieure**. Voici quelques exemples:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 6 \\ 4 & 6 & 5 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ 5 & 2 & 6 \end{bmatrix}.$$

Dans ce cas, la matrice  $\mathbf{A}$  est symétrique, la matrice  $\mathbf{B}$  est diagonale, et la matrice  $\mathbf{C}$  est triangulaire-inférieure.

Une matrice spéciale qu'utilisent fréquemment les économètres est  $\mathbf{I}$ , qui désigne la **matrice identité**. Il s'agit d'une matrice diagonale dont chaque élément diagonal est égal à 1. Un indice est parfois utilisé pour indiquer le nombre de lignes et de colonnes. Ainsi,

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Un vecteur spécial que nous utilisons énormément dans ce livre est  $\mathbf{1}$ , qui désigne un vecteur colonne composé de 1.

La **transposée** d'une matrice est obtenue en échangeant toutes ses écritures lignes et colonnes. Ainsi, le  $ij^{\text{ième}}$  élément de la matrice  $\mathbf{A}$  devient le  $ji^{\text{ième}}$  élément de sa transposée, qui est désignée par la matrice  $\mathbf{A}^\top$ . Notons que certains auteurs utilisent  $\mathbf{A}'$  plutôt que  $\mathbf{A}^\top$  pour désigner la transposée de  $\mathbf{A}$ . La transposée d'une matrice symétrique est égale à la matrice elle-même. La transposée d'un vecteur colonne est un vecteur ligne, et vice versa. Voici quelques exemples:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 5 & 5 \end{bmatrix} \quad \mathbf{A}^\top = \begin{bmatrix} 1 & 3 \\ 2 & 5 \\ 4 & 5 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad \mathbf{b}^\top = [1 \quad 3 \quad 5].$$

L'addition et la soustraction des matrices fonctionnent exactement de la même façon que pour les scalaires, à condition que les matrices puissent être additionnées ou soustraites seulement si elles sont **conformes**. Dans le cas de l'addition et de la soustraction, ceci signifie simplement qu'elles doivent avoir les mêmes dimensions. Si  $\mathbf{A}$  et  $\mathbf{B}$  sont conformes, alors un élément type de  $\mathbf{A} + \mathbf{B}$  est simplement  $A_{ij} + B_{ij}$ , et un élément type de  $\mathbf{A} - \mathbf{B}$  est  $A_{ij} - B_{ij}$ .

En fait, la multiplication matricielle comprend à la fois des additions et des multiplications. Elle est basée sur ce qui est appelé **produit intérieur**, ou **produit scalaire**, de deux vecteurs. Supposons que  $\mathbf{a}$  et  $\mathbf{b}$  soient des vecteurs de dimensions  $n$ . Alors leur produit intérieur est

$$\mathbf{a}^\top \mathbf{b} = \mathbf{b}^\top \mathbf{a} = \sum_{i=1}^n a_i b_i. \quad (\text{A.01})$$

Quand les deux matrices sont multipliées, chaque élément du résultat est égal au produit intérieur d'une des lignes de la première matrice avec une des colonnes de la seconde matrice. Ainsi, si  $\mathbf{C} = \mathbf{AB}$ ,

$$C_{ik} = \sum_{j=1}^m A_{ij} B_{jk}.$$

Ici, nous avons implicitement supposé que la matrice  $\mathbf{A}$  comporte  $m$  colonnes et la matrice  $\mathbf{B}$   $m$  lignes. Pour que les deux matrices soient conformes pour la multiplication, la première matrice doit avoir autant de colonnes que la seconde de lignes. Alors, le résultat a autant de lignes que la première matrice et autant de colonnes que la seconde. Voici un exemple explicite

$$\begin{matrix} \mathbf{A} & \mathbf{B} & = & \mathbf{C} \\ n \times m & m \times l & & n \times l \end{matrix}.$$

Nous voyons rarement ce type de notation dans un livre ou une publication, mais il est souvent commode de l'utiliser lors de calculs destinés à vérifier que les matrices multipliées sont en effet conformes pour définir les dimensions de leur produit.

Le **produit extérieur** des deux vecteurs  $\mathbf{a}$  et  $\mathbf{b}$  est  $\mathbf{ab}^\top$ . Par contraste avec le produit intérieur, qui est un scalaire, le produit extérieur est une matrice de dimension  $n \times n$  si les vecteurs sont de dimension  $n$ .

L'interaction entre la multiplication et l'addition matricielles est intuitive. Il est aisé de vérifier la propriété de **distributivité** à partir des définitions des opérations respectives. Cette propriété est

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}.$$

En plus, ces deux opérations sont **associatives**, ce qui signifie que

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \quad \text{et}$$

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}).$$

La multiplication matricielle est, en général, non commutative. Le fait qu'il soit possible de **prémultiplier** la matrice  $\mathbf{B}$  par la matrice  $\mathbf{A}$  n'implique pas qu'il soit possible de **postmultiplier** la matrice  $\mathbf{B}$  par la matrice  $\mathbf{A}$ . En effet, il est aisé de voir que les deux opérations sont possibles si et seulement si un des produits matriciels est carré; dans ce cas l'autre produit matriciel sera également carré, bien qu'il soit généralement de dimensions différentes. Même quand les deux opérations sont possibles,  $\mathbf{AB} \neq \mathbf{BA}$  sauf dans des cas spéciaux. Les règles pour la multiplication des matrices et des vecteurs sont les mêmes que les règles de multiplication des matrices entre elles; les vecteurs sont simplement traités comme des matrices qui ont une seule colonne ou une seule ligne.

La matrice identité  $\mathbf{I}$  est ainsi appelée parce qu'elle laisse inchangée n'importe quelle matrice avec laquelle elle est soit prémultipliée soit multipliée. Ainsi, pour une matrice quelconque  $\mathbf{A}$ ,  $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ , pourvu naturellement que les deux matrices soient conformes dans chaque cas. Il est facile de voir pourquoi la matrice identité possède cette propriété. Le  $ij^{\text{ième}}$  élément de  $\mathbf{AI}$  est

$$\sum_{k=1}^m A_{ik} \mathbf{I}_{kj} = A_{ij},$$

puisque  $\mathbf{I}_{kj} = 0$  pour  $k \neq j$  et  $\mathbf{I}_{kj} = 1$  pour  $k = j$ . Le vecteur spécial  $\mathbf{e}$  est aussi utile. On l'utilise lorsque l'on désire sommer les éléments d'un autre vecteur, parce que, pour n'importe quel vecteur  $\mathbf{b}$  de dimension  $n$ ,  $\mathbf{e}^\top \mathbf{b} = \sum_{i=1}^n b_i$ .

La transposée du produit de deux matrices est le produit des transposées des matrices en ordre inversé. Ainsi,

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top. \quad (\text{A.02})$$

L'inversion de l'ordre est nécessaire pour que les matrices transposées soient conformes à la multiplication. Le résultat (A.02) peut être prouvé en écrivant les éléments types des deux côtés et en vérifiant qu'ils sont identiques:

$$(\mathbf{AB})_{ij}^\top = (\mathbf{AB})_{ji} = \sum_{k=1}^m A_{jk} B_{ki} = \sum_{k=1}^m (\mathbf{B}^\top)_{ik} (\mathbf{A}^\top)_{kj} = (\mathbf{B}^\top \mathbf{A}^\top)_{ij},$$

où  $m$  est le nombre de colonnes de la matrice  $\mathbf{A}$  et le nombre de lignes de la matrice  $\mathbf{B}$ . Il est toujours possible de multiplier une matrice par sa propre transposée: si la matrice  $\mathbf{A}$  est de dimension  $n \times m$ , alors  $\mathbf{A}^\top$  est de dimension  $m \times n$ , la matrice  $\mathbf{A}^\top \mathbf{A}$  est de dimension  $m \times m$ , et la matrice  $\mathbf{AA}^\top$  est de dimension  $n \times n$ . Ces deux produits matriciels sont symétriques:

$$\mathbf{A}^\top \mathbf{A} = (\mathbf{A}^\top \mathbf{A})^\top \quad \text{et} \quad \mathbf{AA}^\top = (\mathbf{AA}^\top)^\top, \quad (\text{A.03})$$

cela provient directement de l'application de (A.02).

Chaque élément du produit des deux matrices est une somme. Ceci suggère qu'il peut être commode d'utiliser l'algèbre matricielle pour des sommes. Supposons, par exemple, que nous ayons  $n$  observations sur  $k$  régresseurs. Ceux-ci peuvent être arrangés dans une matrice  $\mathbf{X}$  de dimension  $n \times k$ . Ensuite, la matrice des sommes des carrés et des produits croisés des régresseurs peut être écrite de façon compacte comme  $\mathbf{X}^\top \mathbf{X}$ . Il s'agit d'une matrice symétrique de dimension  $k \times k$ , dont un élément diagonal type est  $\sum_{t=1}^n X_{ti}^2$  et un élément non diagonal est  $\sum_{t=1}^n X_{ti} X_{tj}$ .

Il est souvent nécessaire de multiplier une matrice par un scalaire, et ceci fonctionne comme prévu: chaque élément de la matrice est multiplié par le scalaire. De façon occasionnelle, il est nécessaire de multiplier deux matrices, élément par élément. Le résultat est appelé **produit direct** (ou parfois **produit Schur**) des deux matrices. Le produit direct des matrices  $\mathbf{A}$  et  $\mathbf{B}$  est désigné  $\mathbf{A} * \mathbf{B}$ , et un élément type est  $A_{ij} B_{ij}$ .

Une matrice carrée peut ne pas être **inversible**. Si la matrice  $\mathbf{A}$  est inversible, alors elle a une **matrice inverse**  $\mathbf{A}^{-1}$  telle que

$$\mathbf{A} \mathbf{A}^{-1} = \mathbf{A}^{-1} \mathbf{A} = \mathbf{I}.$$

Si la matrice  $\mathbf{A}$  est symétrique, alors la matrice  $\mathbf{A}^{-1}$  l'est aussi. Si la matrice  $\mathbf{A}$  est triangulaire, alors la matrice  $\mathbf{A}^{-1}$  l'est aussi. Sauf dans certains cas spéciaux, il n'est pas facile de calculer l'inverse d'une matrice manuellement. Un tel cas spécial est celui d'une matrice diagonale, disons  $\mathbf{D}$ , avec comme élément type diagonal  $D_{ii}$ . Il est facile de vérifier que  $\mathbf{D}^{-1}$  est aussi une matrice diagonale, avec comme élément type diagonal  $D_{ii}^{-1}$ .

Il est souvent commode d'utiliser la **trace** d'une matrice carrée, qui est simplement la somme des éléments diagonaux. Ainsi,

$$\text{Tr}(\mathbf{A}) = \sum_{i=1}^n A_{ii}.$$

Une propriété très utile est que la trace d'un produit de deux matrices  $\mathbf{A}$  et  $\mathbf{B}$  n'est pas affectée par l'ordre dans lequel les deux matrices sont multipliées. Puisque la trace est définie seulement pour des matrices carrées, à la fois  $\mathbf{AB}$  et  $\mathbf{BA}$  doivent être définies. Ensuite, nous avons

$$\text{Tr}(\mathbf{AB}) = \sum_{i=1}^n (\mathbf{AB})_{ii} = \sum_{i=1}^n \sum_{j=1}^m A_{ij} B_{ji} = \sum_{j=1}^m (\mathbf{BA})_{jj} = \text{Tr}(\mathbf{BA}). \quad (\text{A.04})$$

Le résultat (A.04) peut être développé. Nous considérons un produit (carré) de plusieurs matrices, la trace est invariante à ce qui est appelé **permutation cyclique** des facteurs. Ainsi, par exemple,

$$\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{CAB}) = \text{Tr}(\mathbf{BCA}), \quad (\text{A.05})$$

comme on démontre en appliquant plusieurs fois la relation (A.04). Ce résultat peut être extrêmement commode, et plusieurs résultats standards sur les propriétés des OLS l'utilisent. Par exemple, si  $\mathbf{X}$  est une matrice de dimension  $n \times k$ , (A.05) implique que

$$\text{Tr}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) = \text{Tr}(\mathbf{X}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}) = \text{Tr}(\mathbf{I}_k) = k.$$

### A.3 LA GÉOMÉTRIE DES VECTEURS

Les éléments d'un vecteur de dimension  $n$  peuvent être vus comme les coordonnées d'un point dans un **espace Euclidien** de dimension  $n$ , qui peut être noté  $E^n$ . La différence entre  $E^n$  et l'espace plus familier  $\mathbb{R}^n$  est que le premier inclut une définition spécifique de la **longueur** de chaque vecteur dans  $E^n$ . La longueur d'un vecteur  $\mathbf{x}$  est

$$\|\mathbf{x}\| \equiv (\mathbf{x}^\top \mathbf{x})^{1/2}.$$

Ceci est simplement la racine carrée du produit intérieur de  $\mathbf{x}$  avec lui-même. En termes scalaires, il est simplement

$$\left( \sum_{i=1}^n x_i^2 \right)^{1/2}. \quad (\text{A.06})$$

Comme l'indique la notation  $\|\cdot\|$ , la longueur d'un vecteur est parfois reliée à sa **norme**. Cette définition s'inspire du célèbre théorème de Pythagore concernant les carrés des côtés des triangles rectangles. La définition (A.06) est simplement une généralisation de ce résultat à un nombre arbitraire de dimensions.

Il existe en réalité plus d'une manière de définir un produit intérieur. Celle utilisée auparavant dans (A.01), et la seule utilisée explicitement dans cet ouvrage, est appelée **produit intérieur naturel**. Le produit intérieur naturel de deux vecteurs  $\mathbf{y}$  et  $\mathbf{x}$  est souvent noté  $\langle \mathbf{x}, \mathbf{y} \rangle \equiv \mathbf{x}^\top \mathbf{y}$ . La norme d'un vecteur peut être définie en termes du produit intérieur naturel, puisque  $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$ . L'inégalité fondamentale qui lie des normes et des produits intérieurs est

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (\text{A.07})$$

L'inégalité dans (A.07) devient une égalité si et seulement si  $\mathbf{x}$  et  $\mathbf{y}$  sont **parallèles**, c'est-à-dire si  $\mathbf{y} = \alpha \mathbf{x}$  pour un scalaire  $\alpha$  quelconque.

Le concept de longueur d'un vecteur s'étend naturellement au concept de **distance** entre deux points dans  $E^n$ . Si  $\mathbf{x}, \mathbf{y} \in E^n$ , la distance entre  $\mathbf{x}$  et  $\mathbf{y}$  est  $\|\mathbf{x} - \mathbf{y}\|$ . Notons que cette définition est symétrique par rapport à  $\mathbf{x}$  et  $\mathbf{y}$ . Le concept de produit intérieur nous permet également de définir ce que

nous signifions dans le contexte général par l'**angle** entre deux vecteurs. Pour  $\mathbf{x}, \mathbf{y} \in E^n$ , l'angle  $\phi \equiv \angle(\mathbf{x}, \mathbf{y})$  peut être défini en terme de son **cosinus**,  $\cos \phi$ , comme suit:

$$\cos \phi = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

Cette définition fournit une valeur à  $\cos \phi$  qui varie dans l'intervalle  $[-1, 1]$ , d'après (A.07). La définition est unique seulement si nous limitons la variation possible de  $\phi$  à un intervalle de longueur  $\pi$  (et *non*  $2\pi$ ). De façon habituelle, le meilleur intervalle à choisir est  $[0, \pi]$ . Avec ce choix, l'angle entre un vecteur et lui-même est 0, entre un vecteur et son opposé,  $\pi$ , et entre un vecteur et un autre vecteur qui lui est orthogonal,  $\pi/2$ . Des vecteurs sont **orthogonaux** si leur produit intérieur est nul.

La notion utilisée en économétrie qui correspond le plus étroitement au concept géométrique du cosinus de l'angle est le  $R^2$  d'une régression linéaire. Comme nous l'avons vu dans le Chapitre 1, le  $R^2$  de la régression  $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$  est le carré du cosinus de l'angle entre le vecteur  $\mathbf{y}$  de dimension  $n$  et la projection  $\mathbf{P}_X \mathbf{y}$  de ce vecteur sur l'espace  $\mathcal{S}(\mathbf{X})$  des régresseurs.

Une fois le cosinus de l'angle  $\phi$  trouvé, il est possible de calculer les valeurs de toutes les autres fonctions trigonométriques de  $\phi$ . Ces fonctions sont le **sinus**,  $\sin \phi$ , la **tangente**,  $\tan \phi$ , la **cotangente**,  $\cot \phi$ , la **sécante**,  $\sec \phi$ , et la **cosécante**,  $\csc \phi$ . Parmi celles-ci, la seule qui nous intéresse ici est la cotangente, qui est étroitement reliée aux  $t$  de Student des régressions linéaires. En termes de  $\cos \phi$ ,  $\cot \phi$  est définie comme suit, pour  $\phi \in [0, \pi]$ :

$$\cot \phi = \frac{\cos \phi}{(1 - \cos^2 \phi)^{1/2}}. \quad (\text{A.08})$$

Contrairement au cosinus, qui doit varier entre  $-1$  et  $1$ , la cotangente peut évidemment prendre n'importe quelle valeur réelle.

Pour le cas spécial d'une simple régression linéaire  $\mathbf{y} = \beta \mathbf{x} + \mathbf{u}$  sans terme constant, le  $t$  de Student associé à  $\mathbf{x}$  est

$$\frac{\hat{\beta}}{s(\mathbf{x}^\top \mathbf{x})^{-1/2}}, \quad (\text{A.09})$$

où  $\hat{\beta}$  est l'estimation OLS de  $\beta$ ,  $(\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}$ , et  $s$  est l'estimation OLS de  $\sigma$ , l'écart type des aléas. Dans la notation géométrique, si  $\phi$  est l'angle compris entre  $\mathbf{y}$  et  $\mathbf{x}$ , nous avons

$$\begin{aligned} \hat{\beta} &= \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle} = \frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} \cos \phi, \\ (\mathbf{x}^\top \mathbf{x})^{1/2} &= \|\mathbf{x}\|, \quad \text{et} \\ s^2 &= (n-1)^{-1} (\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}) \\ &= (n-1)^{-1} \|\mathbf{y}\|^2 (1 - \cos^2 \phi). \end{aligned}$$

En substituant ces résultats dans l'expression (A.09) pour le  $t$  de Student, nous trouvons que la valeur de la statistique est

$$(n-1)^{1/2} \frac{\cos \phi}{(1 - \cos^2 \phi)^{1/2}} = (n-1)^{1/2} \cot \phi,$$

d'après (A.08). Consulter le Chapitre 3 pour un résultat analogue dans le contexte de la régression multiple.

## A.4 MATRICES COMME APPLICATIONS DES ESPACES LINÉAIRES

Il est révélateur d'examiner la matrice  $\mathbf{A}$  de dimension  $n \times m$  comme une **application** de  $E^m$  dans  $E^n$ . Cela s'écrit

$$\mathbf{A} : E^m \rightarrow E^n.$$

Notons l'ordre de  $m$  et de  $n$  ici. L'interprétation est simple. Puisque le produit d'une matrice de dimension  $n \times m$  par un vecteur colonne de dimension  $m \times 1$  est défini et fournit un vecteur colonne de dimension  $n \times 1$ , nous pouvons définir l'action de la matrice  $\mathbf{A}$  sur un vecteur  $\mathbf{x}$  de dimension  $m$ ,  $\mathbf{A}(\mathbf{x})$ , comme le produit matriciel  $\mathbf{Ax}$ , et il s'agit d'un vecteur de dimension  $n$ . L'application ainsi définie est linéaire, parce que, si  $\alpha$  et  $\beta$  sont des scalaires quelconques,

$$\mathbf{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{Ax} + \beta\mathbf{Ay},$$

d'après les propriétés classiques des opérations matricielles.

L'espace  $E^m$  des arguments de l'application  $\mathbf{A}$  est appelé **espace de départ** de l'application, et l'espace  $E^n$  des valeurs **espace d'arrivée**. Un sous-espace linéaire important de l'espace de départ est le **noyau** de la matrice. Il est défini comme suit:

$$\mathbf{N}(\mathbf{A}) \equiv \{\mathbf{x} \in E^m \mid \mathbf{Ax} = \mathbf{0}\}.$$

Nous pouvons dire que le noyau de  $\mathbf{A}$  est **annulé** par  $\mathbf{A}$ . Un sous-espace linéaire important de l'espace d'arrivée est appelé **image**, définie par l'expression

$$\mathbf{R}(\mathbf{A}) \equiv \{\mathbf{y} \in E^n \mid \mathbf{y} = \mathbf{Ax} \text{ pour un certain } \mathbf{x} \in E^m\}.$$

L'image peut être décrite comme le sous-espace de  $E^n$  qui contient tous les points **images** d'un point dans  $E^m$  par  $\mathbf{A}$ . L'ensemble des points dans  $E^m$  qui sont appliqués vers un point  $\mathbf{y} \in E^n$ , c'est-à-dire les points qui ont  $\mathbf{y}$  comme image, est appelé **ensemble des antécédents** du point  $\mathbf{y}$ .

Il est clair intuitivement que la **dimension** de l'espace Euclidien  $E^m$  est  $m$ . Nous notons  $\dim E^m = m$ . Quand nous traitons des sous-espaces comme des noyaux ou des images, les dimensions de ces sous-espaces sont moins

apparentes. La nécessaire définition formelle est comme suit. Un espace linéaire est de dimension  $n$  s'il existe  $n$  vecteurs **linéairement indépendants** dans l'espace et si tous les ensembles de plus de  $n$  vecteurs de l'espace sont linéairement dépendants. Un ensemble de vecteurs  $\mathbf{x}_i$ ,  $i = 1, \dots, m$ , est dit **linéairement dépendant** s'il existe une combinaison linéaire non triviale d'entre eux qui est nulle. C'est-à-dire que les  $\mathbf{x}_i$  sont linéairement dépendants s'il existe  $m$  scalaires  $\alpha_i$ , non tous nuls, tels que

$$\sum_{i=1}^m \alpha_i \mathbf{x}_i = \mathbf{0}. \quad (\text{A.10})$$

Pour  $E^m$  lui-même, un ensemble approprié de vecteurs linéairement indépendants est fourni par les vecteurs  $\mathbf{e}_i$ ,  $i = 1, \dots, m$ , de la **base orthonormée** où  $\mathbf{e}_i$  est un vecteur de dimension  $m$  dont le  $i^{\text{ième}}$  élément est 1 et tous les autres sont 0. L'expression du membre de gauche de (A.10), évaluée avec  $\mathbf{e}_i$  à la place de  $\mathbf{x}_i$ , représente le vecteur  $\boldsymbol{\alpha}$  de dimension  $m$  avec comme élément type  $\alpha_i$ . De façon claire, ce vecteur est nul seulement si  $\alpha_i = 0$  pour tout  $i = 1, \dots, m$ , et ainsi les  $\mathbf{e}_i$  sont linéairement indépendants.

Le **complément orthogonal** d'un sous-espace  $\mathbf{M} \subseteq E^m$  est l'espace linéaire

$$\mathbf{M}^\perp \equiv \{ \mathbf{x} \in E^m \mid \mathbf{x}^\top \mathbf{y} = 0 \text{ pour tout } \mathbf{y} \in \mathbf{M} \}.$$

Si  $v$  est la dimension du noyau de la matrice  $\mathbf{A}$  de dimension  $n \times m$  et  $r$  son rang, alors la relation suivante est vraie:

$$m - v = r. \quad (\text{A.11})$$

Ceci signifie que la dimension du complément orthogonal du noyau est égale au rang. Un résultat qui sous-tend toutes les utilisations des matrices de projection au travers de cet ouvrage est que n'importe quel vecteur  $\mathbf{z} \in E^m$  peut être exprimé de manière unique comme la somme de deux vecteurs, l'un dans  $\mathbf{M}$  et l'autre dans  $\mathbf{M}^\perp$ , pour n'importe quel sous-espace de  $E^m$ . Ainsi, nous en déduisons que

$$\dim \mathbf{M} + \dim \mathbf{M}^\perp = m.$$

La dimension de l'image d'une matrice est appelée **rang** de la matrice. Le rang de  $\mathbf{A}$  est parfois noté  $\rho(\mathbf{A})$ . Une matrice  $\mathbf{A}$  de dimension  $n \times m$  est dite de **plein rang** si  $\rho(\mathbf{A})$  est égal au minimum de  $m$  et  $n$ . La terminologie reflète le fait que  $\rho(\mathbf{A})$  ne pourrait jamais excéder  $\min(m, n)$ , comme (A.11) le souligne.

Les  $m$  colonnes d'une matrice de dimension  $n \times m$  peuvent être considérées comme un ensemble de vecteurs de dimension  $n$ . Ainsi, nous pouvons écrire la  $i^{\text{ième}}$  colonne de la matrice  $\mathbf{A}$  comme  $\mathbf{a}_i \in E^n$ . Il est facile de voir que l'image de la matrice  $\mathbf{A}$  est l'ensemble de toutes les combinaisons linéaires de ses colonnes  $\mathbf{a}_i$ . Pour cette raison, l'image de la matrice  $\mathbf{A}$  est souvent appelée



sous-espace engendré par les colonnes de la matrice  $\mathbf{A}$ . Il est commode de noter  $\mathcal{S}(\mathbf{A})$  ce sous-espace, et  $\mathcal{S}^\perp(\mathbf{A})$  son complément orthogonal.

Quand une matrice est interprétée comme une application des espaces linéaires, il est naturel d'attribuer une **norme** à une matrice aussi bien qu'aux vecteurs pour lesquels elle agit. La définition de la norme d'une matrice  $\mathbf{A}$  de dimension  $n \times m$  suit le modèle standard pour la définition des normes des opérateurs. Elle est comme suit:

$$\|\mathbf{A}\| = \max_{\mathbf{x} \in E^m} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}.$$

Il peut être montré que n'importe quelle matrice  $\mathbf{A}$  composée d'éléments finis a une norme finie et que n'importe quelle matrice avec une norme nulle doit simplement être une matrice nulle, c'est-à-dire une matrice dont tous les éléments sont nuls. Si deux matrices  $\mathbf{A}$  et  $\mathbf{B}$  ont des dimensions telles que le produit  $\mathbf{AB}$  existe, alors nous pouvons montrer que

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|.$$

## A.5 MATRICES PARTITIONNÉES

Dans cette section, nous introduisons le concept important d'une **matrice partitionnée** et en dérivons certaines formules très utiles pour l'inversion des matrices partitionnées. Si une matrice  $\mathbf{A}$  possède  $m$  colonnes, et si  $m_1$  et  $m_2$  sont deux entiers positifs tels que  $m_1 + m_2 = m$ , alors nous pouvons définir deux sous-matrices de  $\mathbf{A}$ ,  $\mathbf{A}_1$  et  $\mathbf{A}_2$ , respectivement de dimensions  $n \times m_1$  et  $n \times m_2$ , telles que la sous-matrice  $\mathbf{A}_1$  se compose des  $m_1$  premières colonnes de la matrice  $\mathbf{A}$ , et la sous-matrice  $\mathbf{A}_2$  des  $m_2$  dernières colonnes de la matrice  $\mathbf{A}$ . Nous écrivons

$$\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2]$$

et désignons matrice partitionnée le membre de droite de cette relation.

La partition du cas ci-dessus a été réalisée par colonnes. Nous pouvons également très bien partitionner par lignes ou par lignes et par colonnes, et il peut y avoir plus de deux partitions pour d'autres cas. Les sous-matrices créées par la partition d'une matrice sont appelées les **blocs** de la partition. Si la matrice  $\mathbf{A}$  de dimension  $n \times m$  est partitionnée par ses colonnes et la matrice  $\mathbf{B}$  de dimension  $m \times p$  est partitionnée par ses lignes, la partition peut être conforme. C'est-à-dire que chaque bloc de la partition de la matrice  $\mathbf{A}$  possède autant de colonnes que le bloc correspondant de la partition de la matrice  $\mathbf{B}$  possède de lignes. Dans ce cas, les règles ordinaires de la multiplication matricielle peuvent être appliquées aux matrices partitionnées comme si les blocs étaient réellement les éléments des matrices.

L'utilisation de la partition montre clairement que l'image d'une matrice  $\mathbf{A}$  est l'ensemble de toutes les combinaisons linéaires de ses colonnes  $\mathbf{a}_i$ . Ainsi, partitionnons la matrice  $\mathbf{A}$  de telle sorte que chaque colonne soit traitée comme un bloc:

$$\mathbf{A} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_m].$$

Si la matrice  $\mathbf{A}$  prémultiplie un vecteur  $\mathbf{x}$  de dimension  $m$ , nous pouvons "partitionner"  $\mathbf{x}$  simplement en séparant ses éléments, et obtenons

$$\begin{aligned} \mathbf{Ax} &= [\mathbf{a}_1 \quad \cdots \quad \mathbf{a}_m] \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \\ &= \sum_{i=1}^m \mathbf{a}_i x_i. \end{aligned}$$

Sous cette forme, il est clair que l'image de  $\mathbf{x}$  par la matrice  $\mathbf{A}$  est une combinaison linéaire des colonnes de  $\mathbf{A}$ , définie au moyen des éléments de  $\mathbf{x}$ .

Nous avons remarqué auparavant que les matrices partitionnées peuvent être multipliées si leurs partitions sont conformes, comme si leurs blocs étaient réellement des éléments de matrices. Le résultat d'une telle multiplication partitionnée sera nécessairement une matrice dont la partition en lignes est la même que celle du facteur le plus à gauche du produit matriciel, et dont la partition en colonnes est la même que celle du facteur le plus à droite. Cette propriété peut être utilisée pour démontrer d'autres résultats utiles. Si nous séparons toutes les colonnes du second facteur du produit matriciel  $\mathbf{AB}$ , nous voyons que

$$\mathbf{AB} = \mathbf{A}[\mathbf{b}_1 \quad \cdots \quad \mathbf{b}_m] = [\mathbf{Ab}_1 \quad \cdots \quad \mathbf{Ab}_m],$$

où  $\mathbf{b}_i$  est une colonne type de la matrice  $\mathbf{B}$ . Autrement dit, la  $i^{\text{ième}}$  colonne d'un produit matriciel peut être trouvée en remplaçant le facteur le plus à droite du produit par la  $i^{\text{ième}}$  colonne de ce facteur. De façon similaire, naturellement, la  $i^{\text{ième}}$  ligne d'un produit matriciel est trouvée en remplaçant le facteur le plus à gauche par sa  $i^{\text{ième}}$  ligne.

Supposons que nous considérons une matrice  $\mathbf{X}$  partitionnée en deux groupes de colonnes:  $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$ . La notation est choisie délibérément, parce qu'il est intuitivement utile d'assimiler  $\mathbf{X}$  à une matrice de régresseurs séparés en deux sous-ensembles. En particulier, nous serons capables d'appliquer le Théorème FWL (Section 1.4) dans l'analyse ultérieure. Si la matrice  $\mathbf{X}$  est de dimension  $n \times k$ , alors le produit matriciel  $\mathbf{X}^\top \mathbf{X}$  est de dimension  $k \times k$ . En forme partitionnée, nous avons

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{bmatrix} [\mathbf{X}_1 \quad \mathbf{X}_2] = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{X}_2 \end{bmatrix}. \quad (\text{A.12})$$

Nous allons à présent déduire l'inverse de la matrice partitionnée qui est l'expression la plus à droite dans (A.12). Nous savons que la matrice de covariance des paramètres estimés par OLS pour la régression  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$  est proportionnelle à  $(\mathbf{X}^\top \mathbf{X})^{-1}$ . De plus, si  $\boldsymbol{\beta}$  est partitionnée comme

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix},$$

conformément à la partition de la matrice  $\mathbf{X}$ , alors la matrice de covariance des estimations de  $\boldsymbol{\beta}_1$  est proportionnelle (avec la *même* constante de proportionnalité) à  $(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}$ , où  $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1} \mathbf{X}_2^\top$  est la projection orthogonale sur le complément de l'espace engendré par les colonnes de  $\mathbf{X}_2$ . Ceci signifie que si  $(\mathbf{X}^\top \mathbf{X})^{-1}$  est partitionnée de la même manière que  $\mathbf{X}^\top \mathbf{X}$ , alors le bloc supérieur gauche de l'inverse partitionnée est  $(\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}$ .

Ecrivons  $(\mathbf{X}^\top \mathbf{X})^{-1}$  sous forme partitionnée comme:

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{X}^\top \mathbf{X})_{11}^{-1} & (\mathbf{X}^\top \mathbf{X})_{12}^{-1} \\ (\mathbf{X}^\top \mathbf{X})_{21}^{-1} & (\mathbf{X}^\top \mathbf{X})_{22}^{-1} \end{bmatrix}. \quad (\text{A.13})$$

Nous avons simplement montré que

$$(\mathbf{X}^\top \mathbf{X})_{11}^{-1} = (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}. \quad (\text{A.14})$$

Si (A.12) et (A.13) sont multipliées, le résultat doit être une matrice identité, que nous pouvons partitionner comme

$$\mathbf{I}_k = \begin{bmatrix} \mathbf{I}_{k_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{k_2} \end{bmatrix},$$

où il y a  $k_i$  colonnes dans  $\mathbf{X}_i$  pour  $i = 1, 2$ . Le bloc inférieur gauche de cette matrice identité est  $\mathbf{0}$ , et par une multiplication explicite nous voyons que

$$\mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1} + \mathbf{X}_2^\top \mathbf{X}_2 (\mathbf{X}^\top \mathbf{X})_{21}^{-1} = \mathbf{0},$$

d'où

$$(\mathbf{X}^\top \mathbf{X})_{21}^{-1} = -(\mathbf{X}_2^\top \mathbf{X}_2)^{-1} \mathbf{X}_2^\top \mathbf{X}_1 (\mathbf{X}_1^\top \mathbf{M}_2 \mathbf{X}_1)^{-1}. \quad (\text{A.15})$$

La même sorte de manipulation donnerait une expression pour  $(\mathbf{X}^\top \mathbf{X})_{22}^{-1}$ , mais ceci n'est pas nécessaire, puisque nous savons qu'en inversant les indices 1 et 2 dans l'expression pour  $(\mathbf{X}^\top \mathbf{X})_{11}^{-1}$ ,  $(\mathbf{X}^\top \mathbf{X})_{22}^{-1} = (\mathbf{X}_2^\top \mathbf{M}_1 \mathbf{X}_2)^{-1}$ . Ceci n'est pas l'expression que nous obtiendrions directement, et nous la laissons en exercice pour que le lecteur montre que les deux expressions apparemment différentes sont en fait égales.

Les matrices partitionnées que nous désirons inverser ne sont pas toutes de la forme  $\mathbf{X}^\top \mathbf{X}$ . Nous pouvons obtenir des expressions générales à partir

de ce que nous avons déjà obtenu en écrivant explicitement la matrice de projection orthogonale  $M_2$ . Si la matrice  $X^\top X$  est écrite comme

$$\begin{bmatrix} A & C^\top \\ C & B \end{bmatrix}, \quad (\text{A.16})$$

et la matrice  $(X^\top X)^{-1}$  comme

$$\begin{bmatrix} D & E^\top \\ E & F \end{bmatrix}, \quad (\text{A.17})$$

alors

$$\begin{aligned} D^{-1} &= X_1^\top M_2 X_1 = X_1^\top X_1 - X_1^\top X_2 (X_2^\top X_2)^{-1} X_2^\top X_1 \\ &= A - C^\top B^{-1} C. \end{aligned}$$

Ainsi, de façon très générale, nous avons les relations suivantes entre les blocs des deux matrices inverses partitionnées (A.16) et (A.17):

$$\begin{aligned} D &= (A - C^\top B^{-1} C)^{-1}; \\ E &= -B^{-1} C (A - C^\top B^{-1} C)^{-1} = -(B - C A^{-1} C^\top)^{-1} C A^{-1}; \\ F &= (B - C A^{-1} C^\top)^{-1}. \end{aligned}$$

Ces formules nécessitent que les inverses des blocs diagonaux de la matrice partitionnée d'origine existent.

## A.6 DÉTERMINANTS

Nous avons plusieurs fois fait allusion à la possibilité qu'une matrice carrée puisse ne pas être inversible. Si tel est le cas, alors l'application qui la définit ne sera pas inversible. En général, une application partant d'un espace vers un autre est inversible si et seulement si elle est une **bijection**, ou **bijective**, dans une terminologie mathématique formelle. De façon plus explicite, il faut qu'à chaque point de l'espace d'arrivée de l'application corresponde un et un seul point de l'espace de départ de l'application. Ensuite l'**application inverse**, qui va de l'espace d'arrivée vers l'espace de départ de l'application d'origine, applique chaque point dans l'image vers son unique antécédent.

Nous montrons tout d'abord que seules les matrices carrées sont inversibles. Si  $A$  est une matrice de dimension  $n \times m$ , il est nécessaire pour la rendre inversible que, pour chaque vecteur  $y \in E^n$ , il existe un unique vecteur  $x \in E^m$  tel que  $Ax = y$ . La matrice inverse  $A^{-1}$  est alors une matrice de dimension  $m \times n$  qui transforme un tel vecteur  $y$  en son correspondant  $x$ . Une matrice  $A$  dont le noyau contient plus que le vecteur nul n'est pas inversible. Supposons que  $z \in \mathbf{N}(A)$ ,  $z \neq 0$ ; c'est-à-dire,  $Az = 0$ . Alors, si  $Ax = y$ , nous avons également  $A(x + z) = Ax + Az = Ax$ , et à la fois  $x$  et  $x + z$

doivent appartenir à l'ensemble des antécédents de  $\mathbf{y}$  par  $\mathbf{A}$ , contrairement à la condition d'existence de l'inverse d'une application. Ainsi, si la matrice  $\mathbf{A}$  est de dimension  $n \times m$  et est inversible, nous trouvons à partir de (A.11) que  $m = r$ , la dimension de l'image de  $\mathbf{A}$ . Nous voyons par ailleurs qu'une matrice dont l'image n'est pas le plein espace d'arrivée n'est pas inversible, au quel cas il existe des éléments de celui-ci dont l'ensemble des antécédents est vide, contrairement à la condition pour une inverse. Ceci implique que  $r = n$ , et puisque nous avons déjà vu que  $m = r$ , il s'ensuit que  $m = n$ . Ainsi, nous avons prouvé que seules les matrices carrées sont inversibles. La condition supplémentaire que  $m = r$  implique que seules les matrices carrées de plein rang sont inversibles. Les matrices carrées qui ne sont pas de plein rang sont dites **singulières**, et les matrices carrées de plein rang sont par conséquent parfois dites **non singulières**. Toutes les matrices carrées non singulières sont inversibles.

Comment pouvons-nous savoir si une matrice carrée  $\mathbf{A}$  de dimension  $n \times n$  quelconque est inversible, et si elle l'est, comment peut-on calculer son inverse? Les réponses à ces deux questions sont fournies par le concept du **déterminant** d'une matrice carrée. Puisque, pour le reste de cette section, nous ne traiterons que les matrices carrées, toutes les matrices auxquelles nous ferons référence seront carrées par défaut. Le déterminant d'une matrice est simplement un scalaire. Nous noterons  $|\mathbf{A}|$  le déterminant de la matrice  $\mathbf{A}$  et  $|\det \mathbf{A}|$  désignera la valeur absolue du déterminant de la matrice  $\mathbf{A}$ .

Il est possible de représenter géométriquement le déterminant d'une matrice par le volume de dimension  $n$  de la figure rectiligne générée par les colonnes de la matrice. En deux dimensions, par exemple, les deux colonnes d'une matrice de dimension  $2 \times 2$  définissent un **parallélogramme**, comme cela est montré dans la partie (a) de la Figure A.1. L'aire de ce parallélogramme est le déterminant de la matrice. En trois dimensions, les trois colonnes d'une matrice de dimension  $3 \times 3$  définissent un solide appelé **parallélépipède** (voir la Figure A.2), dont le volume est le déterminant de la matrice. Dans des dimensions supérieures, comme nous le verrons, nous pouvons développer algébriquement le concept du déterminant de manière naturelle, bien qu'il soit évidemment impossible de visualiser les résultats de façon géométrique.

L'aire du parallélogramme est établie dans des textes élémentaires sur la géométrie comme la base fois la hauteur, où la "base" représente la longueur d'un des côtés du parallélogramme, et la "hauteur" la distance *perpendiculaire* entre les deux côtés dont la longueur est la base. Ceci signifie que l'aire d'un parallélogramme peut être calculée comme l'aire d'un rectangle, comme nous l'avons illustré dans la partie (b) de la Figure A.1. De façon algébrique, si les colonnes de la matrice  $\mathbf{A}$  de dimension  $2 \times 2$  sont notées  $\mathbf{a}_1$  et  $\mathbf{a}_2$ , l'aire du parallélogramme est  $\|\mathbf{a}_1\| \|\mathbf{M}_1 \mathbf{a}_2\|$ , où  $\mathbf{M}_1$  est la projection orthogonale sur  $\mathcal{S}^\perp(\mathbf{a}_1)$ . Il est facile de vérifier que nous pouvons échanger les rôles des deux vecteurs sans modifier la valeur de l'aire.

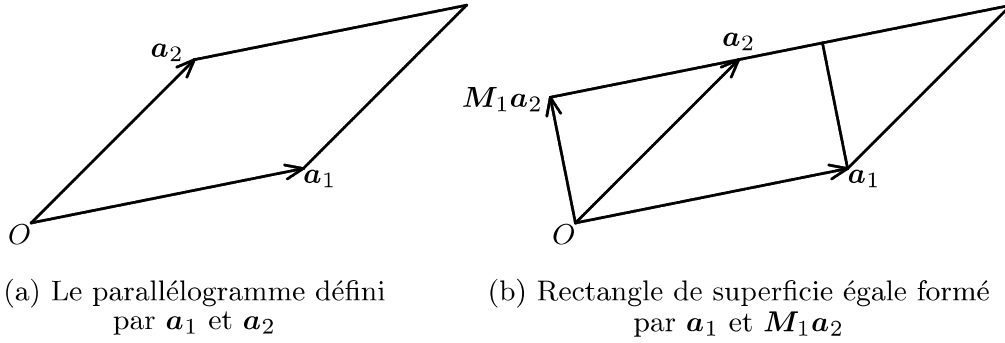


Figure A.1 Déterminants en deux dimensions

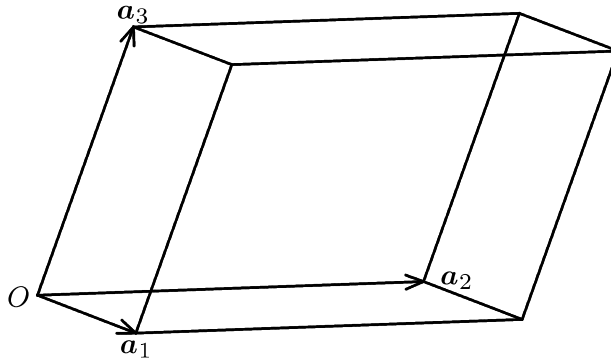


Figure A.2 Un parallélépipède en trois dimensions

Dans le cas à  $n$  dimensions, nous pouvons établir la définition de la valeur absolue du déterminant de la matrice de dimension  $n \times n$   $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_n]$ :

$$\begin{aligned} |\det \mathbf{A}| &= \|M_{(1)}\mathbf{a}_1\| \|M_{(2)}\mathbf{a}_2\| \cdots \|M_{(n-1)}\mathbf{a}_{n-1}\| \|\mathbf{a}_n\| \\ &= \prod_{i=1}^n \|M_{(i)}\mathbf{a}_i\|. \end{aligned} \quad (\text{A.18})$$

Ici,  $M_{(i)}$  la projection orthogonale sur le complément de  $\mathcal{S}(\mathbf{a}_{i+1}, \dots, \mathbf{a}_n)$ , l'espace engendré par les  $n - i$  dernières colonnes de  $\mathbf{A}$ , pour  $i = 1, \dots, n - 1$ . Pour que la seconde ligne soit vraie, il faut que  $M_{(n)} = \mathbf{I}$ .

La définition ci-dessus ne donne que la *valeur absolue* du déterminant. Le signe sera la conséquence d'une autre propriété des déterminants, à savoir, l'anti-symétrie. La valeur de (A.18) est invariante aux changements de l'ordre des colonnes de la matrice  $\mathbf{A}$ , mais quand le signe est pris en compte, nous ferons en sorte qu'une permutation de n'importe laquelle des deux colonnes de la matrice  $\mathbf{A}$  change le signe du déterminant. Considérons la matrice partitionnée suivante:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \mathbf{0} \\ \mathbf{b} & \mathbf{B} \end{bmatrix}. \quad (\text{A.19})$$

Quand la première colonne est projetée sur le complément orthogonal de l'espace engendré par les autres, le résultat sera une colonne avec  $a_{11}$  comme premier élément et des 0 ailleurs. Ainsi, d'après (A.18), la valeur absolue de  $|\mathbf{A}|$  est simplement  $|a_{11}||\det \mathbf{B}|$ . La règle pour le signe du déterminant est une règle récursive: nous supposons que  $|\mathbf{B}|$  a un signe et le multiplions ensuite par celui de l'élément  $a_{11}$  pour obtenir le signe de  $|\mathbf{A}|$ . Pour terminer l'opération, il faut que le signe du déterminant d'une matrice de dimension  $1 \times 1$  soit celui du seul élément de la matrice.

Dans un moment, nous aurons besoin d'utiliser le fait que le déterminant de la matrice (A.19), qui ne dépend pas du vecteur  $\mathbf{b}$  de dimension  $(n-1)$ , est égal au déterminant de n'importe quelle matrice comme (A.19), ayant une colonne nulle à la place de  $\mathbf{b}$  mais avec un vecteur ligne  $\mathbf{c}^\top$  quelconque à la place des éléments nuls dans (A.19). Ainsi, le déterminant de la matrice

$$\begin{bmatrix} a_{11} & \mathbf{c}^\top \\ \mathbf{0} & \mathbf{B} \end{bmatrix} \quad (\text{A.20})$$

est égal à celui de (A.19). Pour comprendre ceci, souvenons-nous que la valeur absolue du déterminant est invariante à l'ordre des colonnes, et sélectionnons la première colonne de (A.20) comme la colonne qui n'est soumise à aucune projection dans (A.18). Toutes les autres colonnes seront alors projetées sur le complément orthogonal de l'espace engendré par la première colonne et perdront par conséquent leurs premiers éléments, c'est-à-dire les éléments de  $\mathbf{c}^\top$ .

Une matrice triangulaire inférieure est un cas particulier de (A.19) dans lequel la matrice  $\mathbf{B}$  est elle-même triangulaire inférieure. De façon similaire, une matrice triangulaire supérieure est un cas particulier de (A.20) dans lequel la matrice  $\mathbf{B}$  est elle-même triangulaire supérieure. Le fait que le déterminant de ces deux matrices soit égal à  $|a_{11}||\det \mathbf{B}|$  implique que si une matrice  $\mathbf{A}$  est triangulaire, son déterminant est égal au produit de ses éléments diagonaux. Pour obtenir ce résultat, nous appliquons simplement le résultat d'origine tout d'abord à  $\mathbf{A}$ , puis à son bloc inférieur droit, enfin au bloc inférieur droit de ce bloc, et ainsi de suite.

Une autre propriété des déterminants est qu'ils sont invariants à des permutations de leurs lignes aussi bien que de leurs colonnes, à un changement de signe près. C'est ce qui ressort de (A.18), puisque la norme d'un vecteur ne dépend pas de la façon dont les lignes sont ordonnées; consulter (A.06).

Le calcul des déterminants n'est évidemment pas une opération linéaire. Ainsi, en général,  $|\mathbf{A} + \mathbf{B}| \neq |\mathbf{A}| + |\mathbf{B}|$ . Cependant, il est vrai que si une colonne d'une matrice est exprimée comme la somme de deux vecteurs, alors le déterminant est additif colonne par colonne. Cela signifie que

$$\begin{aligned} & |\mathbf{a}_1 + \mathbf{b}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n| \\ &= |\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n| + |\mathbf{b}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n|. \end{aligned} \quad (\text{A.21})$$

Ici la notation  $|\cdot|$  avec les blocs d'une matrice partitionnée à l'intérieur désigne le déterminant de la matrice. Pour voir pourquoi (A.21) est vraie, observons que le rang de la projection  $\mathbf{M}_{(2)}$  est seulement 1. Il s'ensuit que, pour n'importe quels vecteurs  $\mathbf{a}$  et  $\mathbf{b}$  de dimension  $n$ ,  $\|\mathbf{M}_{(2)}(\mathbf{a} + \mathbf{b})\| = \|\mathbf{M}_{(2)}\mathbf{a}\| + \|\mathbf{M}_{(2)}\mathbf{b}\|$ . Le résultat provient de ce fait et de la définition (A.18).

Le résultat (A.21) nous permet d'établir la méthode classique d'évaluation manuelle des déterminants. Cette méthode est le **développement du déterminant** par une ligne ou une colonne. Plus personne ne calcule réellement les déterminants de cette manière, sauf peut-être pour le cas trivial  $2 \times 2$ , mais notre discussion sur la façon de développer les déterminants mènera à un certain nombre de résultats utiles. Nous développerons à partir de la première colonne. Pour procéder de la sorte, nous avons besoin d'une notation particulière. Désignons  $\mathbf{A}_{ij}$  la sous-matrice de dimension  $(n-1) \times (n-1)$  de la matrice  $\mathbf{A}$  obtenue en effaçant la  $i^{\text{ième}}$  ligne et la  $j^{\text{ième}}$  colonne. Soit  $A_{ij}$  le déterminant de cette sous-matrice. Nous appelons  $(-1)^{i+j} A_{ij}$  le **cofacteur** de l'élément  $a_{ij}$  dans la matrice  $\mathbf{A}$ . Soit  $\boldsymbol{\alpha}_i$  le vecteur de dimension  $n$  dont tous les éléments sont nuls sauf le  $i^{\text{ième}}$ , qui égale  $a_{i1}$ . Notons alors que les applications successives de (A.21) produisent

$$|\mathbf{A}| = \sum_{i=1}^n |\boldsymbol{\alpha}_i \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n|. \quad (\text{A.22})$$

Si nous écrivons la  $i^{\text{ième}}$  ligne de la somme indicée par  $i$  dans (A.22) comme  $[a_{i1} \quad \mathbf{c}_i^{\top}]$ , alors la  $i^{\text{ième}}$  ligne peut être déplacée pour devenir la première, par un processus de  $i-1$  permutations de lignes, qui génère un facteur de  $(-1)^{i-1}$ . Le résultat est le déterminant

$$(-1)^{i-1} \begin{vmatrix} a_{i1} & \mathbf{c}_i^{\top} \\ \mathbf{0} & \mathbf{A}_{i1} \end{vmatrix},$$

dont la valeur est  $a_{i1} A_{i1}$ , d'après la définition d'un cofacteur. Ainsi, le déterminant (A.22) peut être écrit comme

$$|\mathbf{A}| = \sum_{i=1}^n a_{i1} A_{i1}. \quad (\text{A.23})$$

Puisque  $A_{i1}$  est lui-même un déterminant, (A.23) permet une évaluation récursive d'un déterminant quelconque.

Nous voyons aisément qu'il est possible d'évaluer la matrice  $\mathbf{A}$  en développant par n'importe quelle ligne ou colonne. Formellement,

$$|\mathbf{A}| = \sum_{i=1}^n a_{ij} A_{ij} = \sum_{i=1}^n a_{ji} A_{ji} \quad (\text{A.24})$$



pour tout  $j = 1, \dots, n$ . Ce résultat montre à son tour que  $|\mathbf{A}^\top| = |\mathbf{A}|$ . Si nous développons un déterminant par une colonne, disons la  $j^{\text{ième}}$ , et si nous utilisons de **faux cofacteurs**, c'est-à-dire ceux qui correspondent à une autre colonne, disons la  $k^{\text{ième}}$ , alors nous trouvons que

$$\sum_{i=1}^n a_{ij} A_{ik} = 0. \quad (\text{A.25})$$

Ceci est valable parce que (A.25) est le développement correct du déterminant d'une matrice dans laquelle la  $k^{\text{ième}}$  colonne est remplacée par la  $j^{\text{ième}}$  colonne. N'importe quelle matrice dans laquelle au moins deux colonnes sont identiques a un déterminant nul, puisque quand la même colonne survient une seconde fois dans (A.18), elle sera projetée sur le complément orthogonal de l'espace qu'elle engendre, en donnant un vecteur de norme nulle.

Pour la même raison, n'importe quelle matrice dans laquelle une colonne est une combinaison linéaire des autres aura un déterminant nul. Une matrice qui satisfait cette condition ne sera pas de plein rang, et nous voyons aussi qu'une matrice singulière a nécessairement un déterminant nul. Il n'est pas difficile de voir que la réciproque est vraie: une matrice avec un déterminant nul est nécessairement singulière. Tout ceci est également pertinent de façon géométrique, naturellement. Si une matrice de dimension  $n \times n$  n'est pas de plein rang, le parallélépipède défini par la matrice sera un objet de dimension inférieure à  $n$ , et ainsi son volume (dans l'espace de dimension  $n$ ) sera nul.

Les résultats (A.24) et (A.25) peuvent être utilisés pour construire l'inverse d'une matrice non singulière  $\mathbf{A}$ . Considérons la matrice  $\mathbf{B}$  avec comme élément type  $b_{ij} \equiv A_{ji}$ , qui est juste la transposée de la matrice des cofacteurs. Nous voyons que

$$(\mathbf{AB})_{ij} = \sum_{k=1}^n a_{ik} A_{jk} = |\mathbf{A}| \delta_{ij},$$

où  $\delta_{ij}$  est le delta de Kronecker, égal à 1 si  $i = j$  et à 0 sinon. Ainsi,  $\mathbf{AB} = |\mathbf{A}| \mathbf{I}$ , de sorte que  $|\mathbf{A}|^{-1} \mathbf{B}$ , qui existe si et seulement si  $|\mathbf{A}| \neq 0$ , doit être l'inverse de  $\mathbf{A}$ .

Le résultat (A.24) nous permet de calculer les *dérivées partielles* du déterminant d'une matrice par rapport aux éléments de la matrice. Le cofacteur  $A_{ij}$  est le déterminant d'une matrice qui ne contient aucun élément de la  $i^{\text{ième}}$  ligne ou de la  $j^{\text{ième}}$  colonne de la matrice  $\mathbf{A}$ . Il s'ensuit que la dérivée partielle de  $|\mathbf{A}|$  par rapport à  $a_{ij}$  est juste  $A_{ij}$ , qui est  $|\mathbf{A}|$  fois le  $ji^{\text{ième}}$  élément de la matrice  $\mathbf{A}^{-1}$ . Ce résultat peut être écrit avec la notation matricielle comme

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}| (\mathbf{A}^{-1})^\top.$$

A partir de ce dernier, nous pouvons en déduire le résultat encore plus utile selon lequel

$$\frac{\partial \log |\mathbf{A}|}{\partial \mathbf{A}} = (\mathbf{A}^{-1})^\top.$$

Bien que le déterminant d'une somme de matrices ne soit pas en général la somme des déterminants, le déterminant d'un produit de matrices *est* le produit des déterminants. Soit  $\mathbf{A}$  et  $\mathbf{B}$  deux matrices de dimensions  $n \times n$ , toutes deux avec des déterminants non nuls. Ensuite,  $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$ . Un corollaire utile est que  $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ . Ceci provient des propriétés  $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$  et  $|\mathbf{I}| = 1$ .

Pour conclure cette section, nous prouvons un résultat utilisé dans les Chapitres 18 et 20. Selon ce résultat, nous avons

$$\begin{vmatrix} \mathbf{A}^\top \mathbf{A} & \mathbf{A}^\top \mathbf{B} \\ \mathbf{B}^\top \mathbf{A} & \mathbf{B}^\top \mathbf{B} \end{vmatrix} = |\mathbf{A}^\top \mathbf{M}_B \mathbf{A}| |\mathbf{B}^\top \mathbf{B}| = |\mathbf{B}^\top \mathbf{M}_A \mathbf{B}| |\mathbf{A}^\top \mathbf{A}|, \quad (\text{A.26})$$

où  $\mathbf{M}_A$  et  $\mathbf{M}_B$  sont les projections orthogonales des colonnes des matrices  $\mathbf{A}$  et  $\mathbf{B}$ , que l'on peut supposer être de plein rang sans perte de généralité. Nous utilisons les résultats (A.14) et (A.15) sur l'inversion des matrices partitionnées comme précédemment pour écrire

$$\begin{bmatrix} \mathbf{A}^\top \mathbf{A} & \mathbf{A}^\top \mathbf{B} \\ \mathbf{B}^\top \mathbf{A} & \mathbf{B}^\top \mathbf{B} \end{bmatrix} \begin{bmatrix} (\mathbf{A}^\top \mathbf{M}_B \mathbf{A})^{-1} & \mathbf{0} \\ -(\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{A} (\mathbf{A}^\top \mathbf{M}_B \mathbf{A})^{-1} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{A}^\top \mathbf{B} \\ \mathbf{0} & \mathbf{B}^\top \mathbf{B} \end{bmatrix}.$$

Il est évident que le déterminant de la matrice du membre de droite est juste  $|\mathbf{B}^\top \mathbf{B}|$ , tandis que celui du second facteur dans le membre de gauche est  $|\mathbf{A}^\top \mathbf{M}_B \mathbf{A}|^{-1}$ . La première égalité dans (A.26) en découle. La seconde égalité peut être prouvée par un argument similaire, mais en utilisant différentes expressions pour l'inverse de la matrice partitionnée.

## A.7 MATRICES DÉFINIES POSITIVES

Une matrice symétrique  $\mathbf{A}$  de dimension  $n \times n$  est dite **définie positive** si la forme quadratique  $\mathbf{x}^\top \mathbf{A} \mathbf{x}$  est positive pour tout vecteur non nul  $\mathbf{x}$  de dimension  $n$ . Si la forme quadratique peut prendre des valeurs nulles, elle est **semi-définie positive** ou **définie non négative**. Des matrices qui sont **définies négatives** ou **semi-définies négatives** sont définies de façon analogue.

N'importe quelle matrice de la forme  $\mathbf{B}^\top \mathbf{B}$  est définie positive si le rang de la matrice  $\mathbf{B}$  est égal au nombre de colonnes et semi-définie positive sinon. Pour s'en rendre compte, observons que  $\mathbf{B}^\top \mathbf{B}$  est symétrique et que, pour n'importe quel vecteur  $\mathbf{x}$  non nul,

$$\mathbf{x}^\top \mathbf{B}^\top \mathbf{B} \mathbf{x} = (\mathbf{B} \mathbf{x})^\top (\mathbf{B} \mathbf{x}) = \|\mathbf{B} \mathbf{x}\|^2 \geq 0.$$

Ce résultat est valable avec l'égalité à condition que  $\mathbf{B}\mathbf{x} = \mathbf{0}$ . Mais, dans ce cas,  $\mathbf{B}$  ne peut pas être de plein rang, puisque  $\mathbf{B}\mathbf{x} = \mathbf{0}$  signifie que les colonnes de  $\mathbf{B}$  ne sont pas linéairement indépendantes. Un raisonnement similaire montre que si la matrice  $\mathbf{A}$  est définie positive, alors n'importe quelle matrice de la forme  $\mathbf{B}^\top \mathbf{A} \mathbf{B}$  est définie positive si la matrice  $\mathbf{B}$  vérifie la même condition de rang, et semi-définie positive sinon.

Une matrice définie positive ne peut pas être singulière, puisque si la matrice  $\mathbf{A}$  est singulière, il doit exister un vecteur  $\mathbf{x}$  non nul tel que  $\mathbf{A}\mathbf{x} = \mathbf{0}$ . Ce qui implique également que  $\mathbf{x}^\top \mathbf{A} \mathbf{x} = 0$ . Cela signifie que la matrice  $\mathbf{A}$  n'est pas définie positive. Ainsi, l'inverse d'une matrice définie positive existe toujours. Elle est également définie positive, parce que, pour n'importe quel vecteur  $\mathbf{x}$  non nul,

$$\mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x} = \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{A} \mathbf{A}^{-1} \mathbf{x} = (\mathbf{A}^{-1} \mathbf{x})^\top \mathbf{A} (\mathbf{A}^{-1} \mathbf{x}) > 0.$$

Ici l'inégalité provient directement du fait que la matrice  $\mathbf{A}$  est définie positive.

Pour n'importe quelle matrice définie positive  $\mathbf{A}$ , nous pouvons trouver une matrice  $\mathbf{B}$  telle que  $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$ . Il est souvent nécessaire d'élaborer une telle matrice  $\mathbf{B}$  à partir d'une matrice donnée  $\mathbf{A}$  dans des applications économétriques; un exemple est la matrice  $\boldsymbol{\eta}$  définie dans (9.08). Fréquemment, nous souhaitons aller plus loin et trouver une matrice *triangulaire*  $\mathbf{B}$ . Nous esquissons à présent un algorithme pour une telle **décomposition triangulaire**. Il produit une matrice  $\mathbf{B}$  triangulaire supérieure à partir d'une matrice définie positive donnée  $\mathbf{A}$ . Un algorithme analogue pour produire une matrice  $\mathbf{B}$  triangulaire inférieure peut aussi être trouvé.

Nous commençons par définir  $b_{11} = \sqrt{a_{11}}$ , où  $a_{ij}$  et  $b_{ij}$  désignent les  $ij$ <sup>èmes</sup> éléments des matrices  $\mathbf{A}$  et  $\mathbf{B}$ , respectivement. La première ligne entière de la matrice  $\mathbf{B}$  est ainsi obtenue par une application séquentielle de la formule suivante, pour  $j = 2, \dots, n$ :

$$b_{1j} = \frac{a_{1j}}{b_{11}}.$$

Les lignes suivantes sont calculées de façon séquentielle, de telle manière que, au cours du calcul de la  $i$ <sup>ème</sup> ligne, les éléments de la première ligne à la  $(i-1)$ <sup>ème</sup> soient disponibles. Pour la  $i$ <sup>ème</sup> ligne, les éléments  $b_{ij}$  sont initialisés à zéro pour  $j < i$ , puisque la matrice  $\mathbf{B}$  doit être triangulaire supérieure. Alors, le  $i$ <sup>ème</sup> élément diagonal est

$$b_{ii} = \left( a_{ii} - \sum_{k=1}^{i-1} b_{ki}^2 \right)^{1/2}, \quad (\text{A.27})$$

dans lequel le membre entier de droite est connu. Pour compléter la ligne, les éléments  $b_{ij}$  pour  $j > i$  sont déterminés par la formule

$$b_{ij} = \frac{1}{b_{ii}} \left( a_{ij} - \sum_{k=1}^{i-1} b_{ki} b_{kj} \right).$$

A nouveau, tout ce qui apparaît dans le membre de droite est disponible à chaque fois que cela est nécessaire. Un calcul que nous ne reproduirons pas montre que la grandeur dont la racine carrée est calculée dans (A.27) est positive à condition que la matrice  $\mathbf{A}$  soit définie positive, et montre aussi que la matrice  $\mathbf{B}$  générée par l'algorithme satisfait la contrainte  $\mathbf{B}^\top \mathbf{B} = \mathbf{A}$ . Les résultats de la section précédente montrent que le déterminant d'une matrice triangulaire est juste le produit de ses éléments diagonaux. Ainsi, nous pouvons obtenir le déterminant de la matrice  $\mathbf{B}$  presque comme un sous-produit de l'algorithme destiné à trouver la matrice  $\mathbf{B}$ . Le carré du déterminant de la matrice  $\mathbf{B}$  est alors le déterminant de la matrice  $\mathbf{A}$ .

Dans certaines manipulations des matrices de covariance dans le texte, nous utilisons le fait que si  $\mathbf{A}$  et  $\mathbf{B}$  sont des matrices semi-définies positives, alors  $\mathbf{A} - \mathbf{B}$  est une matrice définie positive si et seulement si  $\mathbf{B}^{-1} - \mathbf{A}^{-1}$  l'est. Nous démontrons maintenant ce résultat très utile. Soit  $\mathbf{A}^{-1/2}$  une matrice telle que  $(\mathbf{A}^{-1/2})^\top \mathbf{A}^{-1/2} = \mathbf{A}^{-1}$ . Il peut être vu que

$$\mathbf{A}^{-1/2} \mathbf{A} (\mathbf{A}^{-1/2})^\top = (\mathbf{A}^{-1/2})^\top \mathbf{A} \mathbf{A}^{-1/2} = \mathbf{I}.$$

Tout d'abord nous montrons que si  $\mathbf{I} - \mathbf{A}$  est une matrice définie positive, alors  $\mathbf{A}^{-1} - \mathbf{I}$  l'est également et réciproquement. Ceci provient du résultat, prouvé auparavant, qu'en prémultipliant une matrice définie positive par n'importe quelle matrice de plein rang et en multipliant ensuite le résultat par la transposée de cette matrice, nous obtenons une matrice définie positive. Ainsi, la caractéristique définie positive de  $\mathbf{I} - \mathbf{A}$  implique celui de  $(\mathbf{A}^{-1/2})^\top (\mathbf{I} - \mathbf{A}) \mathbf{A}^{-1/2}$ , qui est juste  $\mathbf{A}^{-1} - \mathbf{I}$ . Le résultat réciproque provient de l'inversion des positions des matrices  $\mathbf{A}$  et  $\mathbf{A}^{-1}$ .

Si  $\mathbf{A} - \mathbf{B}$  est définie positive, alors  $\mathbf{A}^{-1/2} (\mathbf{A} - \mathbf{B}) (\mathbf{A}^{-1/2})^\top$  l'est également, c'est-à-dire  $\mathbf{I} - \mathbf{A}^{-1/2} \mathbf{B} (\mathbf{A}^{-1/2})^\top$ . Le caractère défini positif de cette dernière matrice entraîne que de  $(\mathbf{A}^{1/2})^\top \mathbf{B}^{-1} \mathbf{A}^{1/2} - \mathbf{I}$ , où la matrice  $\mathbf{A}^{1/2}$  est l'inverse de la matrice  $\mathbf{A}^{-1/2}$ , et également de  $(\mathbf{A}^{-1/2})^\top (\mathbf{A}^{1/2})^\top \mathbf{B}^{-1} \mathbf{A}^{1/2} \mathbf{A}^{-1/2} - (\mathbf{A}^{-1/2})^\top \mathbf{A}^{-1/2}$ , qui est juste  $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ , comme requis. A nouveau, le résultat réciproque provient de l'inversion des positions des matrices et de leurs inverses. Un résultat similaire est vrai pour les matrices semi-définies positives:  $\mathbf{A} - \mathbf{B}$  est une matrice semi-définie positive si et seulement si  $\mathbf{B}^{-1} - \mathbf{A}^{-1}$  l'est.

## A.8 VALEURS PROPRES ET VECTEURS PROPRES

Un scalaire  $\lambda$  est une **valeur propre** d'une matrice  $\mathbf{A}$  s'il existe un vecteur non nul  $\mathbf{x}$  tel que

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \quad (\text{A.28})$$

Ainsi, l'action de la matrice  $\mathbf{A}$  sur  $\mathbf{x}$  produit un vecteur de même direction que  $\mathbf{x}$ , mais de longueur différente à moins que  $\lambda = 1$ . Le vecteur  $\mathbf{x}$  est appelé

le **vecteur propre** qui correspond à la valeur propre  $\lambda$ . Bien que ces idées soient définies de façon très générale, nous restreindrons notre attention ici aux valeurs propres et vecteurs propres des matrices symétriques réelles.

La relation des valeurs propres (A.28) implique que

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}, \quad (\text{A.29})$$

à partir de laquelle nous concluons que la matrice  $\mathbf{A} - \lambda \mathbf{I}$  est singulière. Son déterminant,  $|\mathbf{A} - \lambda \mathbf{I}|$  est par conséquent égal à zéro. Il peut être montré de différentes façons que ce déterminant est un polynôme en  $\lambda$ , de degré  $n$  si la matrice  $\mathbf{A}$  est de dimension  $n \times n$ . Le théorème fondamental de l'algèbre nous indique qu'un tel polynôme possède  $n$  racines complexes, disons  $\lambda_1, \dots, \lambda_n$ . A chaque  $\lambda_i$  doit correspondre un vecteur propre  $\mathbf{x}_i$ . Ce vecteur propre est déterminé à un facteur d'échelle près, parce que si  $\mathbf{x}_i$  est un vecteur propre qui correspond à  $\lambda_i$ , alors  $\alpha \mathbf{x}_i$  l'est également pour n'importe quel scalaire  $\alpha$ . Le vecteur propre  $\mathbf{x}_i$  n'a pas nécessairement des éléments réels si  $\lambda_i$  elle-même n'est pas réelle.

Si  $\mathbf{A}$  est une matrice réelle symétrique, nous pouvons montrer que les valeurs propres  $\lambda_i$  sont en fait toutes réelles et qu'il est également possible de choisir des vecteurs propres réels. Si  $\mathbf{A}$  est une matrice définie positive, alors toutes ses valeurs propres sont positives. Ceci provient des faits que

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \lambda \mathbf{x}^\top \mathbf{x}$$

et qu'à la fois  $\mathbf{x}^\top \mathbf{x}$  et  $\mathbf{x}^\top \mathbf{A} \mathbf{x}$  sont positives. Les vecteurs propres d'une matrice symétrique réelle peuvent être choisis comme mutuellement orthogonaux. Si nous regardons les deux vecteurs propres  $\mathbf{x}_i$  et  $\mathbf{x}_j$ , qui correspondent aux deux valeurs propres distinctes  $\lambda_i$  et  $\lambda_j$ , alors  $\mathbf{x}_i$  et  $\mathbf{x}_j$  sont nécessairement orthogonaux:

$$\lambda_i \mathbf{x}_j^\top \mathbf{x}_i = \mathbf{x}_j^\top \mathbf{A} \mathbf{x}_i = (\mathbf{A} \mathbf{x}_j)^\top \mathbf{x}_i = \lambda_j \mathbf{x}_j^\top \mathbf{x}_i,$$

ce qui est impossible à moins que  $\mathbf{x}_j^\top \mathbf{x}_i = 0$ . Si toutes les valeurs propres ne sont pas distinctes, alors deux (ou plusieurs) vecteurs propres peuvent correspondre à une seule et même valeur propre. Quand cela survient, ces deux vecteurs propres engendrent un espace qui est orthogonal à toutes les autres valeurs propres d'après le raisonnement précédemment établi. Puisque n'importe quelle combinaison linéaire des deux vecteurs propres sera également un vecteur propre qui correspond à la valeur propre, nous pouvons choisir un ensemble orthogonal de ceux-ci. Ainsi, que les valeurs propres soient toutes distinctes ou non, nous pouvons sélectionner des vecteurs propres **orthonormaux**, c'est-à-dire des vecteurs mutuellement orthogonaux et normés à 1. Ainsi, les vecteurs propres d'une matrice symétrique réelle fournissent une base orthonormée.

Soit  $\mathbf{U} \equiv [\mathbf{x}_1 \ \cdots \ \mathbf{x}_n]$  une matrice dont les colonnes sont un ensemble orthogonal de vecteurs propres de la matrice  $\mathbf{A}$ , qui correspondent aux valeurs

propres  $\lambda_i$ ,  $i = 1, \dots, n$ . Nous pouvons alors résumer en une seule relation l'ensemble des relations (A.28) entre valeurs propres et vecteurs propres:

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \quad (\text{A.30})$$

où  $\mathbf{\Lambda}$  est une matrice diagonale avec  $\lambda_i$  pour  $i^{\text{ième}}$  élément diagonal. La  $i^{\text{ième}}$  colonne de  $\mathbf{A}\mathbf{U}$  est  $\mathbf{A}\mathbf{x}_i$ , et la  $i^{\text{ième}}$  colonne de  $\mathbf{U}\mathbf{\Lambda}$  est  $\lambda_i\mathbf{x}_i$ . Puisque les colonnes de  $\mathbf{U}$  sont orthonormales, nous trouvons que  $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ , qui implique que  $\mathbf{U}^\top = \mathbf{U}^{-1}$ . Une telle matrice  $\mathbf{U}$  est dite **matrice orthogonale**. La postmultiplication de (A.30) par  $\mathbf{U}^\top$  fournit

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top. \quad (\text{A.31})$$

Cette équation exprime la **diagonalisation** de la matrice  $\mathbf{A}$ .

Le calcul des déterminants des deux côtés de (A.31) fournit

$$|\mathbf{A}| = |\mathbf{U}||\mathbf{U}^\top||\mathbf{\Lambda}| = |\mathbf{U}||\mathbf{U}^{-1}||\mathbf{\Lambda}| = |\mathbf{\Lambda}| = \prod_{i=1}^n \lambda_i,$$

calcul à partir duquel nous déduisons le résultat important que le déterminant d'une matrice est le produit de ses valeurs propres. En fait, ce résultat est également valable pour les matrices non symétriques.

Un résultat utilisé dans le Chapitre 18 est que si la matrice  $\mathbf{A}$  est définie positive et si la matrice  $\mathbf{B}$  semi-définie positive, alors

$$|\mathbf{A} + \mathbf{B}| \geq |\mathbf{A}|.$$

Nous montrons ceci tout d'abord pour le cas particulier où  $\mathbf{A} = \mathbf{I}$  puis en déduisons le résultat général. L'équation qui définit les valeurs propres de la matrice  $\mathbf{I} + \mathbf{B}$  est

$$|\mathbf{I} + \mathbf{B} - \lambda\mathbf{I}| = 0,$$

à partir de (A.29). Ceci devient

$$|\mathbf{B} - (\lambda - 1)\mathbf{I}| = 0.$$

Il s'ensuit que les valeurs propres  $\lambda_i$  de  $\mathbf{I} + \mathbf{B}$  satisfont l'équation  $\lambda_i = 1 + \mu_i$ , où  $\mu_i$  est une valeur propre de la matrice  $\mathbf{B}$ . Si  $\mathbf{B}$  est une matrice semi-définie positive, ses valeurs propres sont toutes supérieures ou égales à 0, ce qui implique que les valeurs propres de la matrice  $\mathbf{I} + \mathbf{B}$  sont toutes supérieures ou égales à 1. Puisque le déterminant d'une matrice est le produit de ses valeurs propres, nous concluons que le déterminant de la matrice  $\mathbf{I} + \mathbf{B}$  est supérieur ou égal à 1, valeur du déterminant de la matrice  $\mathbf{I}$ .

Soit  $\mathbf{A}^{1/2}$  une matrice telle que  $\mathbf{A}^{1/2}(\mathbf{A}^{1/2})^\top = \mathbf{A}$ . Alors, si  $\mathbf{B}$  est une matrice semi-définie positive,

$$\begin{aligned} |\mathbf{A} + \mathbf{B}| &= |\mathbf{A}^{1/2}(\mathbf{I} + \mathbf{A}^{-1/2}\mathbf{B}(\mathbf{A}^{-1/2})^\top)(\mathbf{A}^{1/2})^\top| \\ (\text{A.32}) \quad &= |(\mathbf{A}^{1/2})|^2 |\mathbf{I} + \mathbf{A}^{-1/2}\mathbf{B}(\mathbf{A}^{-1/2})^\top|. \end{aligned}$$

La matrice  $\mathbf{A}^{-1/2}\mathbf{B}(\mathbf{A}^{-1/2})^\top$  est semi-définie positive parce que la matrice  $\mathbf{B}$  l'est, ce qui rend le dernier facteur dans (A.32) supérieur à 1. Puisque

$$|(\mathbf{A}^{1/2})|^2 = |\mathbf{A}|,$$

nous voyons que  $|\mathbf{A} + \mathbf{B}| \geq |\mathbf{A}|$ , comme prévu.

## TERMES ET CONCEPTS

angle entre deux vecteurs	matrice inversible
application définie par une matrice	matrice orthogonale
application inverse	matrice partitionnée
base orthogonale	matrice semi-définie négative
base orthonormée	matrice semi-définie-positive (ou définie non négative)
bijection	matrice symétrique
blocs d'une matrice partitionnée	matrice triangulaire
cofacteur	matrice triangulaire inférieure
complément orthogonal (d'un sous-espace)	matrice triangulaire supérieure
décomposition triangulaire	matrices conformes
déterminant	norme (d'une matrice)
développement du déterminant (d'après une ligne ou une colonne)	noyau (d'une matrice)
diagonale principale d'une matrice carrée	parallélépipède
diagonalisation (d'une matrice symétrique réelle)	parallélogramme
dimension (d'un espace Euclidien)	permutation cyclique (des facteurs d'un produit de matrice)
distance entre deux points dans $E^n$	plein rang
espace d'arrivée d'une application	postmultiplication
espace de départ d'une application	prémultiplication
espace engendré (des colonnes d'une matrice)	produit direct (produit Schur)
espace euclidien de dimension $n$ , $E^n$	produit extérieur
faux cofacteurs	produit intérieur naturel
fonctions trigonométriques: sinus, cosinus, tangente, cotangente, sécante, cosécante	produit intérieur (produit scalaire)
image d'une matrice	propriété associative (pour l'addition et la multiplication matricielle)
image et antécédent	propriété distributive (pour l'addition et la multiplication matricielle)
longueur (ou norme) d'un vecteur	rang d'une matrice
matrice carrée	trace d'une matrice
matrice carrée non singulière	transposée d'une matrice
matrice carrée singulière	valeur propre
matrice définie négative	vecteur colonne
matrice définie positive	vecteur ligne
matrice diagonale	vecteur propre
matrice identité	vecteurs linéairement dépendants
matrice inverse	vecteurs linéairement indépendants
	vecteurs orthogonaux
	vecteurs parallèles



# Annexe B

## Résultats de la Théorie des Probabilités

### B.1 INTRODUCTION

Les lecteurs de cet ouvrage devraient déjà être relativement familiers avec la théorie des probabilités et la statistique. Cette annexe a été élaborée pour aider ceux qui souhaitent rafraîchir leur mémoire et pour réunir les résultats pour faciliter les références. Il ne s'agit en aucun cas d'un substitut à des manuels de second cycle tels que ceux Casella et Berger (1990) ou Spanos (1986). La Section B.2 rappelle les concepts de base des variables aléatoires et des distributions de probabilité. La Section B.3 traite des moments des variables aléatoires et de certains résultats connexes. Enfin la Section B.4 fait le point sur certaines des distributions de probabilité les plus communément utilisées en économétrie.

### B.2 VARIABLES ALÉATOIRES ET LOIS DE PROBABILITÉ

Le concept de **variable aléatoire** sous-tend la majeure partie de la théorie des probabilités et de sa discipline affiliée de la statistique. Une définition complètement formelle d'une variable aléatoire nécessite le concept d'**espace probabilisable**, sur lequel peut se définir une **sigma-algèbre**, qui sert à son tour de support à la définition d'une **mesure de probabilité**. Nous ne pouvons pas dans cet ouvrage détailler tous ces concepts, et les lecteurs intéressés sont orientés vers Billingsley (1979) pour un traitement approprié.

L'essentiel de nos propos, très simplifié, s'expose comme suit. La première composante nécessaire est un ensemble d'éléments que nous appellerions communément les "états du monde" dans la théorie économique traditionnelle. Cet ensemble plus formellement appelé **espace des événements** ou **espace des réalisations**, peut être très simple. Par exemple, si nous traitions le lancé d'une pièce de monnaie, il serait composé de deux éléments, pile ou face. Dans d'autres circonstances, il peut être très compliqué afin de pouvoir gérer tous les détails d'un processus stochastique à indice soit discret, comme les suites de variables aléatoires rencontrées dans la théorie asymptotique exposée dans cet ouvrage, soit continu. Un exemple de cette dernière possibilité est celui du processus de Wiener mentionné dans le Chapitre 20. Dans tous les cas, l'espace des réalisations doit posséder une structure suffisamment riche

pour que chaque réalisation *possible* soit représentée par un point de l'espace; des réalisations différentes doivent correspondre à des points différents.

Bien que chaque réalisation possible doive être représentée dans l'espace des réalisations, il n'est pas toujours possible d'attribuer une probabilité à toutes ces réalisations. Même si c'était le cas, la probabilité associée pourrait ne pas être particulièrement riche en information. Par exemple, si nous considérons une seule variable aléatoire pouvant prendre n'importe quelle valeur sur la droite réelle, la probabilité qu'elle prenne un nombre réel particulier est traditionnellement nulle. Des probabilités positives seraient dans ce cas associées uniquement à des intervalles de longueur positive. Une structure est par conséquent nécessaire pour déterminer précisément quels sont les sous-ensembles de l'espace des réalisations — les **événements composites** dans la terminologie probabiliste standard — auxquels nous allons attribuer des probabilités. Cette structure est la sigma-algèbre dans la théorie formelle.

La dernière composante essentielle est la mesure de probabilité: la manière dont les probabilités sont effectivement attribuées à des événements, composites ou simples. La seule chose à conserver à l'esprit ici est que les mesures de probabilité doivent respecter les lois de probabilité dictées par notre intuition. Ces lois sont remarquablement simples: la probabilité de l'événement nul (rien ne se réalise) est nulle, la probabilité de l'espace entier des réalisations (une réalisation quelconque se produit) est égale à un, et la probabilité qu'un ensemble quelconque d'événements disjoints, ou qui s'excluent mutuellement, se réalise est égale à la somme des probabilités de chacun des événements disjoints pris séparément.

Nous pouvons à présent livrer une définition non formelle de ce que nous entendons par variable aléatoire, ou **v.a.** en abrégé. Le cas le plus simple est celui d'une **variable aléatoire scalaire**, qui ne prend qu'une seule valeur réelle. Une telle variable aléatoire sera une application de l'espace des réalisations dans la droite réelle, c'est-à-dire l'attribution d'un nombre réel à chaque réalisation possible. Un instant de réflexion nous montrera que c'est précisément ce que nous entendons par variable aléatoire: une grandeur dont la valeur prise dépend de l'état du monde. En général, une application quelconque de l'espace des réalisations dans la droite réelle n'est pas à proprement parler une variable aléatoire, parce que nous insistons sur le fait qu'il devrait être possible de définir une **distribution de probabilité** pour chaque variable aléatoire. Le sens de ceci, plus spécifiquement, est que, si  $x$  est une v.a. quelconque, nous devrions être capables d'attribuer des probabilités à des événements tels que  $(x \leq X)$  pour tout réel  $X$ . Notons  $\Omega$  l'espace des réalisations; c'est une notation très répandue dans la théorie des probabilités. Alors l'événement  $(x \leq X)$  peut être explicité sous la forme du sous-ensemble suivant de  $\Omega$ :

$$(\omega \in \Omega \mid x(\omega) \leq X). \quad (\text{B.01})$$

Le fait que  $x$  soit une *application* de  $\Omega$  dans la droite donne son sens à (B.01).

Pour que  $x$  soit une variable aléatoire bien définie, il doit être possible d'attribuer une probabilité à chacun des ensembles (B.01). Cela nous conduit à la **fonction de densité cumulée**, ou **c.d.f.**, ou **fonction de distribution** ou encore **fonction de répartition** de la variable aléatoire  $x$ , que l'on note souvent  $F(x)$  et qui est définie sur la droite réelle. Du fait que la valeur d'une c.d.f. est une probabilité, une c.d.f. doit prendre ses valeurs dans l'intervalle  $[0, 1]$ . Une c.d.f. type est définie par une équation de la forme

$$F_x(X) = \Pr(\omega \in \Omega \mid x(\omega) \leq X).$$

Habituellement, il est pratique d'omettre la référence à  $\omega$  et  $\Omega$  en écrivant simplement  $\Pr(x \leq X)$ . Par construction, une c.d.f. tend vers zéro lorsque son argument tend vers  $-\infty$ , et vers un lorsque son argument tend vers  $+\infty$ . De plus, ce doit être une fonction faiblement croissante en son argument. Cette propriété est vraie parce que, si  $X_1 < X_2$ , alors l'événement  $(x \leq X_1)$  est compris dans l'événement  $(x \leq X_2)$  et ne peut donc avoir une probabilité supérieure à celle de  $(x \leq X_2)$ . Montrer ce résultat en détail à l'aide de la règle sur la somme des probabilités d'ensembles d'événements disjoints est un bon exercice.

Les variables aléatoires peuvent prendre des valeurs sous forme de vecteurs, de matrices, ou bien d'autres formes encore. Une variable aléatoire qui prend des valeurs sous la forme d'un vecteur est appelée **variable aléatoire vectorielle**. Les propriétés probabilistes d'une v.a. vectorielle  $\mathbf{x}$  peuvent être représentées par une généralisation de la c.d.f. appelée **c.d.f. jointe**. Si  $\mathbf{x} \in \mathbb{R}^n$ , alors sa c.d.f. est une fonction de  $n$  arguments, comme suit:

$$F_{\mathbf{x}}(X_1, \dots, X_n) = \Pr((x_1 \leq X_1) \cap \dots \cap (x_n \leq X_n)).$$

Ici  $x_i$  désigne la  $i^{\text{ième}}$  composante de  $\mathbf{x}$ , et le symbole  $\cap$  désigne l'intersection d'ensembles: l'événement en question est l'ensemble de tous les  $\omega \in \Omega$  tel que  $x_1 \leq X_1$  et  $x_2 \leq X_2$ , et ainsi de suite. Une c.d.f. jointe possède des propriétés similaires à la c.d.f. d'une variable aléatoire scalaire. Elle tend vers zéro quand *n'importe lequel* de ses arguments tend vers  $-\infty$ , et vers un lorsque *tous* ses arguments tendent vers  $+\infty$ . A partir d'une c.d.f. jointe, nous pouvons dériver la **distribution marginale** de n'importe quelle composante de  $\mathbf{x}$ . Cela correspond simplement à la probabilité d'une composante considérée comme une variable aléatoire scalaire. Cette distribution marginale est bien sûr représentée par une c.d.f. ordinaire, qui pour la composante  $x_i$  est obtenue en initialisant à  $+\infty$  toutes les composantes de la c.d.f. jointe autres que  $x_i$ :

$$F_{x_i}(X_i) = F_{\mathbf{x}}(+\infty, \dots, X_i, \dots, +\infty).$$

Ceci est alors la probabilité que  $x_i \leq X_i$  et que toutes les composantes de  $\mathbf{x}$  autres que  $x_i$  prennent n'importe quelle valeur. La distribution marginale de tout sous-ensemble de composantes  $\mathbf{x}$  est représentée de manière analogue

par une c.d.f. jointe provenant des c.d.f. d'origine en initialisant à  $+\infty$  toutes les composantes non sélectionnées.

Les distributions de probabilité jointe permettent d'introduire la notion importante d'**indépendance statistique**. Soit  $\mathbf{x}$  une variable aléatoire vectorielle de dimension  $n$ , et supposons qu'elle est partitionnée comme  $\mathbf{x} = [\mathbf{x}_1 : \mathbf{x}_2]$ , avec  $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ ,  $\mathbf{x}_2 \in \mathbb{R}^{n_2}$ , et  $n_1 + n_2 = n$ . Alors  $\mathbf{x}_1$  et  $\mathbf{x}_2$  sont dits **statistiquement indépendants**, ou souvent plus simplement indépendants, si la c.d.f. jointe du vecteur  $\mathbf{x}$  est le produit des c.d.f. de  $\mathbf{x}_1$  et  $\mathbf{x}_2$ . Dans une notation simplifiée, cela signifie que

$$F_{\mathbf{x}}(\mathbf{X}_1, \mathbf{X}_2) = F_{\mathbf{x}}(\mathbf{X}_1, \infty_2) F_{\mathbf{x}}(\infty_1, \mathbf{X}_2),$$

où  $\infty_1$  et  $\infty_2$  désignent les vecteurs dont les composantes sont égales à  $+\infty$ .

Le concept de **fonction de densité de probabilité**, ou **p.d.f.**, est très étroitement relié à celui de c.d.f. Bien qu'une fonction de distribution existe pour toute variable aléatoire bien définie, une p.d.f. n'existe que si la c.d.f. est *différentiable*. Pour une v.a. scalaire, la fonction de densité, souvent notée  $f$ , est simplement la dérivée de la c.d.f.:

$$f_x(X) \equiv F'_x(X).$$

La **densité jointe** d'un ensemble de v.a., ou de manière équivalente une v.a. vectorielle, s'obtient en dérivant la c.d.f. jointe par rapport à *tous* ses arguments:

$$f_{\mathbf{x}}(X_1, \dots, X_n) = \frac{\partial^n F_{\mathbf{x}}(X_1, \dots, X_n)}{\partial X_1 \cdots \partial X_n}.$$

Le fait qu'une c.d.f. varie de 0 à 1 implique que la fonction de densité soit **normalisée** pour que son intégrale soit égale à un. En effet,

$$\begin{aligned} \int_{-\infty}^{\infty} f_x(X) dX &= \int_{-\infty}^{\infty} F'_x(X) dX \\ &= [F_x(X)]_{X=-\infty}^{X=+\infty} = 1 - 0 = 1. \end{aligned} \tag{B.02}$$

De la même manière nous montrons que l'intégrale multiple d'une fonction de densité jointe par rapport à ses arguments lorsqu'ils varient de  $-\infty$  à  $+\infty$  est égale à un. Un résultat encore plus utile est que l'intégrale d'une p.d.f. jointe par rapport à certains arguments seulement fournit la densité marginale des variables par rapport auxquelles on n'a pas "intégré". Celle-ci est appelée **densité marginale**. Si deux groupes de v.a. sont indépendants, alors il est aisé de voir que l'indépendance en terme des c.d.f. implique que la densité jointe des deux groupes est le produit des densités marginales de ces deux groupes.

Une autre propriété cruciale d'une fonction de densité est qu'elle est non négative. Cela provient directement de sa définition de dérivée d'une fonction faiblement croissante. Mais c'est également le pendant d'une propriété

très utile d'une densité, qui nous permet de l'utiliser pour calculer les probabilités d'événements associés à une variable aléatoire donnée. Supposons que  $x$  soit une v.a. scalaire. Alors pour tout intervalle  $[a, b]$  de la droite réelle, nous pourrions souhaiter calculer la probabilité que  $x \in [a, b]$ . Cela provient directement de la définition d'une c.d.f. que, si  $a < b$ ,

$$\Pr(x \in [a, b]) = F_x(b) - F_x(a).$$

Par un argument similaire à celui conduisant à (B.02), cette probabilité est

$$\int_a^b f_x(X) dX. \quad (\text{B.03})$$

Puisque (B.03) doit être vraie pour des valeurs quelconques de  $a$  et  $b$ , il est clair que  $f_x$  doit être une fonction non négative.

### B.3 MOMENTS DES VARIABLES ALÉATOIRES

L'une des propriétés les plus importantes que peut posséder une variable aléatoire est une **espérance**. La définition de l'espérance d'une v.a. scalaire suffira; pour des v.a. vectorielles ou matricielles, les espérances sont définies élément par élément. Ainsi, si  $x$  est une matrice aléatoire scalaire, son espérance est définie comme la valeur de l'intégrale

$$\int_{-\infty}^{\infty} X dF_x(X), \quad (\text{B.04})$$

si elle existe. Le type d'intégrale dans (B.04) est appelé **intégrale de Stieltjes**, en raison de la présence de la **fonction d'intégration**  $F_x$ . Les lecteurs pour qui le concept d'une intégrale de Stieltjes est nouveau peuvent souhaiter consulter un article standard sur l'analyse réelle, tel que celui de Burrill et Knudsen (1969) ou celui de Mukherjea et Pothoven (1984), pour les détails. Nous ne les fournirons pas ici, parce qu'ils ne sont pas très importants pour l'analyse que nous livrons. La principale caractéristique d'une intégrale de Stieltjes, en ce qui nous concerne, est que si la fonction d'intégration est dérivable, il est possible d'exprimer l'intégrale de Stieltjes comme une intégrale ordinaire en terme de sa dérivée. Pour (B.04), nous obtenons l'expression suivante pour l'espérance de  $x$ :

$$\int_{-\infty}^{\infty} X f_x(X) dX, \quad (\text{B.05})$$

où  $f_x$  est la densité de  $x$ . Pour simplifier notre discussion ultérieure, nous ne traiterons que des c.d.f. dérivables.

Toutes les variables aléatoires ne possèdent pas une espérance. L'intégrale d'une fonction de densité doit toujours exister et être égale à 1. Mais, puisque

$X$  va de  $-\infty$  à  $\infty$ , l'intégrale (B.05) peut diverger vers une des limites d'intégration, ou les deux, si la densité  $f_x$  ne tend pas vers zéro suffisamment rapidement. Par un léger abus de terminologie, l'espérance d'une variable aléatoire est parfois appelée sa **moyenne**. A proprement parler, une moyenne est une propriété d'un *échantillon* de réalisations de v.a., plutôt que d'une distribution de probabilité. Dans les rares circonstances où la confusion est possible, l'espérance peut être appelée **moyenne de la population** pour la distinguer de la **moyenne d'échantillon**.

On fait souvent référence à l'espérance d'une variable aléatoire en tant que son **moment d'ordre un**. Les **moments** dits **d'ordre supérieur** s'ils existent, sont les espérances des puissances de la v.a. Ainsi le **moment d'ordre deux** d'une variable aléatoire  $x$  est l'espérance de  $x^2$ , le **moment d'ordre trois** l'espérance de  $x^3$ , et ainsi de suite. Les moments non entiers sont définis de manière analogue, mais nous ne les utiliserons pas dans cet ouvrage. En général, le moment d'ordre  $k$  de la v.a.  $x$  est

$$m_k \equiv \int_{-\infty}^{\infty} X^k f_x(X) dX.$$

Observons que la valeur de tout moment ne dépend que de la distribution de probabilité de la v.a. en cause. Pour cette raison, on parle souvent des moments de la distribution plutôt que de ceux d'une variable aléatoire particulière. Notons également que si une distribution possède un moment d'ordre  $k$ , elle possède également tous les moments d'ordre inférieur à  $k$ .

La définition précédente concerne les **moments non centrés** d'une distribution. Il est probablement plus ordinaire de travailler avec les **moments centrés**, définis comme les moments ordinaires de la différence entre la variable aléatoire et son espérance. Ainsi, si  $E(x)$  représente l'espérance de  $x$ , le moment centré d'ordre  $k$  de la distribution de  $x$  est

$$\bar{m}_k \equiv E(x - E(x))^k.$$

Le moment centré le plus important est de loin le moment d'ordre deux. C'est la **variance** de la v.a. La notation usuelle pour une variance est  $\sigma^2$ , et cette notation souligne le fait qu'une variance ne peut pas être négative. La racine carrée,  $\sigma$ , est appelée **écart standard** de la distribution. Les *estimations* des écarts standards sont souvent appelées **écarts types**, en particulier lorsque la variable aléatoire en cause est un paramètre estimé.

Il est souvent important de pouvoir définir les moments de v.a. vectorielles. Pour le moment d'ordre un, c'est trivial: le moment d'ordre un d'une variable aléatoire vectorielle  $\mathbf{x}$  de dimension  $n$  est simplement le vecteur ordinaire  $\bar{\mathbf{x}}$  de dimension  $n$  dont l'élément type est  $\bar{x}_i \equiv E(x_i)$ . Pour les moments d'ordres deux et supérieurs, cela se complique. Pour les moments centrés d'ordre deux, il est nécessaire de définir une matrice de dimension  $n \times n$ , parfois appelée **matrice de variance**, parfois **matrice de covariance**, et parfois

**matrice de variance-covariance.** La terminologie n'est pas standard, et nous préférons l'expression du milieu. La matrice de covariance de  $\mathbf{x}$  sera notée  $\mathbf{V}(\mathbf{x})$  et définie par

$$\mathbf{V}(\mathbf{x}) \equiv E((\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^\top).$$

Les éléments diagonaux de  $\mathbf{V}(\mathbf{x})$  sont les variances des éléments de  $\mathbf{x}$ . L'élément non diagonal  $V_{ij}$  est appelé **covariance** de  $x_i$  et  $x_j$ . Les moments d'ordre supérieur de v.a. vectorielles peuvent être définis de manière analogue. Ils font appel à des objets possédant plus de deux indices et ne sont pas utilisés dans cet ouvrage.

Si nous calculons l'espérance du produit de deux variables aléatoires indépendantes, le résultat correspond simplement au produit des espérances des variables aléatoires prises séparément. Cela provient du fait que la densité jointe de deux v.a. indépendantes est le produit des deux densités marginales. De plus, la covariance de deux variables aléatoires indépendantes est nulle. Une question embarrassante standard en théorie des probabilités consiste à savoir si deux v.a. de covariance nulle sont nécessairement indépendantes: la réponse est “non”. Cependant, une covariance nulle *est* une condition suffisante pour que l'espérance du produit de deux variables aléatoires soit égale au produit des espérances séparées.

Il est souvent nécessaire de calculer la variance d'une combinaison linéaire de variables aléatoires. Supposons que ces v.a. soient les éléments de la v.a. vectorielle  $\mathbf{x}$ , et que la combinaison linéaire d'intérêt s'écrive  $\mathbf{a}^\top \mathbf{x}$  pour un vecteur non aléatoire quelconque  $\mathbf{a}$ . Il est facile de montrer que la variance de cette combinaison linéaire est  $\mathbf{a}^\top \mathbf{V}(\mathbf{x}) \mathbf{a}$ . De manière similaire, si l'on construit un vecteur de combinaisons linéaires des éléments de  $\mathbf{x}$ , en construisant par exemple  $\mathbf{A}^\top \mathbf{x}$  pour une matrice non aléatoire quelconque  $\mathbf{A}$  adéquate, alors

$$\mathbf{V}(\mathbf{A}^\top \mathbf{x}) = \mathbf{A}^\top \mathbf{V}(\mathbf{x}) \mathbf{A}. \quad (\text{B.06})$$

Si une variable aléatoire possède une variance, sa valeur peut être utilisée pour calculer une borne pour la masse de probabilité contenue dans la queue de distribution. Nous entendons par **queue** d'une distribution de probabilité un événement de la forme  $(x > X)$  ou  $(x < X)$ , où  $X$  est substantiellement à la droite du centre de la distribution dans le premier cas et substantiellement à gauche dans le second. Le premier cas définit la **queue de droite** de la distribution et le second la **queue de gauche**. Le terme ambigu “centre” est ici employé du fait que la définition même de queue est imprécise. Nous pourrions entendre par centre l'espérance, la médiane, le mode ou toute autre **mesure de tendance centrale**. L'imprécision provient sûrement du fait que les v.a. n'ont pas toutes une espérance. Pour une v.a. ne possédant pas d'espérance, différentes mesures de tendance centrale peuvent être appropriées. Parfois, c'est la probabilité qu'une variable aléatoire appartienne à une queue de distribution qui nous intéresse, parfois c'est la probabilité qu'elle appartienne à

la queue de droite, et parfois qu'elle appartienne à la queue de gauche. Les queues de gauche sont d'un intérêt très limité lorsque la v.a. ne prend que des valeurs positives.

La borne sur la masse de probabilité dans les queues à laquelle nous avons fait allusion est connue sous le nom d'**inégalité de Chebyshev**. Nous pouvons la dériver comme suit. Supposons que le moment non centré d'ordre deux de la v.a.  $x$  soit  $V$ . Si  $x$  est elle-même une variable aléatoire centrée alors  $E(x) = 0$  et  $V$  est sa variance. L'inégalité de Chebyshev établit que, pour tout nombre positif  $\alpha$ ,

$$\Pr(|x| > \alpha) \leq \frac{V}{\alpha^2}. \quad (\text{B.07})$$

Pour le comprendre, notons que la définition de  $V$  est

$$V = E(x^2) = \int_{-\infty}^{\infty} X^2 f_x(X) dX.$$

Cette intégrale peut se décomposer en une somme de trois intégrales:

$$V = \int_{-\alpha}^{\alpha} X^2 f_x(X) dX + \int_{\alpha}^{\infty} X^2 f_x(X) dX + \int_{-\infty}^{-\alpha} X^2 f_x(X) dX. \quad (\text{B.08})$$

Considérons les deux derniers termes du membre de droite. Le facteur  $X^2$  dans l'intégrande est toujours supérieur à  $\alpha^2$  sur le domaine d'intégration de ces termes. Ainsi ces termes sont au moins supérieurs à

$$\alpha^2 \left( \int_{\alpha}^{\infty} f_x(X) dX + \int_{-\infty}^{-\alpha} f_x(X) dX \right) = \alpha^2 \Pr(|x| > \alpha),$$

grâce à (B.03). Puisque tous les termes de (B.08) sont non négatifs, nous concluons que

$$V \geq \alpha^2 \Pr(|x| > \alpha).$$

La réorganisation de cette inégalité conduit à (B.07). De là découle une forme plus familière de l'inégalité de Chebyshev, qui établit que, pour une variable aléatoire  $x$  d'espérance  $\mu$  et de variance  $\sigma^2$ ,

$$\Pr\left(\left|\frac{x - \mu}{\sigma}\right| > \alpha\right) \leq \frac{1}{\alpha^2}.$$

Le calcul de l'espérance d'une variable aléatoire est une opération linéaire. Si  $x$  et  $y$  sont deux v.a. et  $a$  et  $b$  deux réels non aléatoires, alors  $E(ax + by) = aE(x) + bE(y)$ . Cela provient directement de la définition (B.05) d'une espérance. En général, cependant, si  $g$  est une fonction scalaire d'une variable aléatoire scalaire  $x$ , nous *n'avons pas*  $E(g(x)) = g(E(x))$ . Cette conclusion serait vraie uniquement si  $g$  était une **fonction affine**, ce qui signifie que  $g(x) = ax + b$  pour deux réels  $a$  et  $b$ .



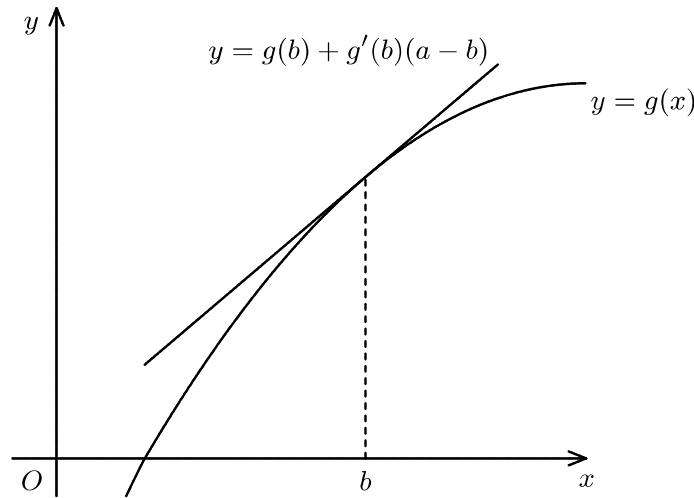


Figure B.1 Une fonction concave type

Par ailleurs, si la fonction  $g$  est concave ou convexe, nous pouvons montrer que l'inégalité entre  $E(g(x))$  et  $g(E(x))$  a un signe particulier. Ce résultat est connu sous le nom d'**inégalité de Jensen**. Pour être concrets et parce que c'est un cas qui survient dans la théorie du maximum de vraisemblance du Chapitre 8, supposons que  $g$  soit une fonction concave comme la fonction logarithmique. Alors l'inégalité assure que

$$E(g(x)) \leq g(E(x)).$$

Pour le comprendre, supposons que  $g$  soit dérivable, bien que le résultat ne nécessite pas cette hypothèse. Alors une manière d'exprimer la concavité de  $g$  est l'inégalité

$$g(a) \leq g(b) + g'(b)(a - b), \quad \text{pour tous réels } a, b. \quad (\text{B.09})$$

Cette inégalité est illustrée sur la Figure B.1, qui devrait donner l'intuition de l'inégalité de Jensen autant que (B.09) elle-même. Notons  $\bar{x}$  l'espérance  $E(x)$ . Alors

$$\begin{aligned} E(g(x)) &= \int_{-\infty}^{\infty} g(X) f_x(X) dX \\ &\leq \int_{-\infty}^{\infty} (g(\bar{x}) + g'(\bar{x})(X - \bar{x})) f_x(X) dX, \end{aligned}$$

où l'inégalité provient de (B.09). La seconde ligne est ici égale à

$$\begin{aligned} &g(\bar{x}) + g'(\bar{x}) \left( \int_{-\infty}^{\infty} X f_x(X) dX - \bar{x} \int_{-\infty}^{\infty} f_x(X) dX \right) \\ &= g(E(x)) + g'(\bar{x})(\bar{x} - \bar{x}) = g(E(x)). \end{aligned}$$

Ceci démontre alors l'inégalité de Jensen pour le cas dérivable.

Si n'importe quelle fonction d'une variable aléatoire  $x$  est évaluée en  $x$ , le résultat est une autre variable aléatoire. Ceci vaut aussi bien pour la fonction de densité  $f_x$  que pour toute autre fonction. En économétrie, on est rarement intéressé par une seule fonction de densité mais davantage par une famille paramétrique de fonctions de densité. Dans le cas simple où il n'existe qu'un seul paramètre, une telle famille peut s'écrire  $f(x, \theta)$ , où  $\theta$  est le paramètre. Le logarithme de cette fonction est la **fonction de logvraisemblance** associée à la famille paramétrique. Une propriété importante d'une telle fonction de logvraisemblance est que, sous des conditions de régularité appropriées, la dérivée de  $\log f(x, \theta)$  est une variable aléatoire telle que, si sa moyenne est calculée avec la densité correspondant à la même valeur de  $\theta$  que celle utilisée pour évaluer la dérivée, cette moyenne est nulle si elle existe. Il est utile d'esquisser une démonstration de ce résultat, qui peut s'exprimer comme

$$E_\theta \left( \frac{\partial \log f}{\partial \theta} \right) = 0, \quad (\text{B.10})$$

où l'indice  $\theta$  de l'opérateur d'espérance indique que celle-ci est calculée avec  $f(\cdot, \theta)$ .

La démonstration de (B.10) utilise un résultat standard sur la dérivation des intégrales. Ce résultat établit que la dérivée d'une intégrale de la forme

$$\int_{a(\theta)}^{b(\theta)} g(y, \theta) dy$$

par rapport au paramètre  $\theta$  peut s'exprimer en terme des dérivées des fonctions  $a$ ,  $b$ , et  $g$  par rapport à  $\theta$ , à condition qu'elles existent, et est égale à

$$-a'(\theta)g(a(\theta), \theta) + b'(\theta)g(b(\theta), \theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial g(y, \theta)}{\partial \theta} dy,$$

à nouveau à condition que l'intégrale du dernier terme existe. Pour ce résultat standard, consulter n'importe quel manuel sur l'analyse réelle, tels que Burrill et Knudsen (1969) ou Mukherjea et Pothoven (1984).

Pour démontrer (B.10), nous tirons profit du fait que la fonction de densité  $f$  est normalisée pour que son intégrale soit égale à un pour toutes les valeurs du paramètre  $\theta$ . Supposons que le **support** de la fonction de densité soit l'intervalle  $[a(\theta), b(\theta)]$  pour tout  $\theta$ . Cela signifie que la densité est nulle en dehors de cet intervalle ou que la probabilité qu'une v.a. distribuée avec la densité  $f(\cdot, \theta)$  prenne une valeur en dehors de cet intervalle est nulle. Alors la condition de normalisation est

$$\int_{a(\theta)}^{b(\theta)} f(y, \theta) dy = 1.$$

Puisque cette condition est valable pour toutes les valeurs admissibles de  $\theta$ , nous pouvons la dériver par rapport à  $\theta$  et obtenir

$$-a'(\theta)f(a(\theta)) + b'(\theta)f(b(\theta)) + \int_{a(\theta)}^{b(\theta)} \frac{\partial f(y, \theta)}{\partial \theta} dy = 0. \quad (\text{B.11})$$

Le dernier terme, l'intégrale, peut s'exprimer comme

$$\int_{a(\theta)}^{b(\theta)} f(y, \theta) \frac{\partial \log f(y, \theta)}{\partial \theta} dy = E_{\theta} \left( \frac{\partial \log f}{\partial \theta} \right).$$

Nous voyons que, hormis les conditions de régularité sur la dérivabilité et l'existence de l'espérance de  $\partial \log f / \partial \theta$ , le résultat (B.10) requiert que les deux premiers termes dans (B.11) s'annulent pour une raison ou pour une autre. Une condition évidente menant à ce résultat est que les bornes du support de la fonction de densité soient indépendantes du paramètre  $\theta$ . Par exemple, si l'espace d'intégration est la droite réelle, nous aurons automatiquement le résultat voulu. Une condition différente est que la densité s'annule aux bornes du support, et cela arrive en fait fréquemment dans la pratique. Des difficultés peuvent malgré tout survenir si le support dépend de  $\theta$  et si la densité n'est pas nulle sur ses bornes.

Le raisonnement utilisé pour établir (B.10) peut être employé aussi bien pour établir l'égalité de la matrice d'information de la théorie du maximum de vraisemblance; voir le Chapitre 8.

## B.4 QUELQUES DISTRIBUTIONS DE PROBABILITÉ USUELLES

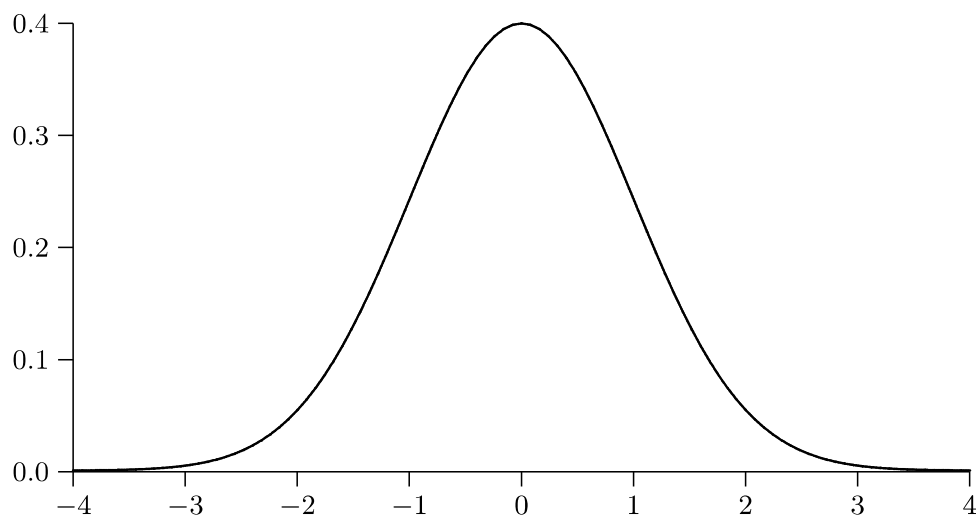
La distribution de probabilité la plus importante est sans conteste la **distribution normale centrée réduite**. Cette distribution apparaît très souvent dans la théorie économétrique, et les définitions d'un grand nombre d'autres distributions communément employées utilisent directement la distribution normale centrée réduite. La distribution normale possède la densité dont le tracé est la plus ou moins célèbre courbe en cloche des ouvrages d'initiation à la statistique, et elle représente parfois la distribution des notes d'examen; voir la Figure B.2.

La densité de la distribution normale centrée réduite est définie sur la droite réelle comme suit:

$$\phi(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2\right). \quad (\text{B.12})$$

Contrairement à cette p.d.f., qui s'exprime uniquement en terme de fonctions standards, la c.d.f. de la distribution normale centrée réduite doit être définie explicitement comme l'intégrale

$$\Phi(x) = \int_{-\infty}^x \phi(y) dy.$$



**Figure B.2** La densité de la loi normale centrée réduite

Remarquons que  $\phi$  et  $\Phi$  sont les notations traditionnelles des p.d.f. et c.d.f. de la distribution normale centrée réduite. Bien que  $\Phi$  ne puisse pas s'exprimer avec des fonctions standards, il est facile de l'évaluer numériquement.<sup>1</sup> Il est aisé de vérifier que  $\phi$  satisfait toutes les exigences pour une densité de probabilité: partout positive, intégrale égale à l'unité. Par conséquent, puisque  $\Phi$  est définie en terme d'une densité adéquate, elle doit satisfaire les exigences pour une c.d.f.

Du fait de la symétrie de la densité (B.12) par rapport à zéro, l'espérance de la densité normale est nulle, tout comme le sont les moments d'ordre impair de la distribution. Les moments d'ordre pair ne sont pas difficiles à calculer. La variance peut se calculer à l'aide d'une intégration par parties. Puisque la dérivée de  $\phi(x)$  est  $-x\phi(x)$ , l'intégrale indéfinie de  $x\phi(x)$  est  $-\phi(x)$ . Par conséquent,

$$\begin{aligned} \int_{-\infty}^{\infty} x^2 \phi(x) dx &= \int_{-\infty}^{\infty} x (x\phi(x)) dx \\ &= [-x\phi(x)]_{x=-\infty}^{x=\infty} + \int_{-\infty}^{\infty} \phi(x) dx = 1, \end{aligned} \quad (\text{B.13})$$

et nous voyons que la variance d'une loi normale centrée réduite est égale à un. Cette propriété justifie l'usage du terme "réduite" dans ce contexte. Les moments d'ordre pair supérieur de la densité de la normale centrée réduite sont quasiment aussi faciles à calculer. Le résultat, obtenu par un calcul de

<sup>1</sup> Notons que, dans chacune des définitions précédentes, nous avons par souci de simplicité abandonné l'usage des variables en majuscules. Il ne devrait subsister aucun risque de confusion entre des variables ordinaires et des variables aléatoires dans ce qui suit.

réurrence d'une intégration par parties comparable à celle dans (B.13), est que

$$m_{2k} = (2k-1)(2k-3) \cdots (3)(1).$$

Ainsi le moment d'ordre 4 est  $(3)(1) = 3$ , le moment d'ordre 6 est  $(5)(3)(1) = 15$ , et ainsi de suite.

Toute v.a. normalement distribuée d'espérance non nulle et de variance non unitaire peut se définir par une translation et une normalisation d'une variable normale centrée réduite. La famille des distributions ainsi définie doit posséder deux paramètres que l'on peut noter  $\mu$ , l'espérance, et  $\sigma^2$ , la variance. Si  $y$  est distribuée normalement avec une espérance  $\mu$  et une variance  $\sigma^2$ , nous disons qu'elle a une **distribution normale univariée**. Nous écrivons  $y \sim N(\mu, \sigma^2)$ . La densité de  $y$  est

$$\frac{1}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) = (2\pi)^{-1/2} \frac{1}{\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \quad (\text{B.14})$$

Nous pouvons dériver ce résultat de (B.12) à l'aide d'un résultat sur les transformations des variables aléatoires que nous démontrerons dans un instant. Si  $y \sim N(\mu, \sigma^2)$ , alors nous montrons que la v.a.  $x \equiv (y - \mu)/\sigma$  possède une espérance nulle et une variance unitaire. De fait,  $x \sim N(0, 1)$ , ce qui correspond à la manière traditionnelle d'écrire la distribution normale centrée réduite.

Une extension importante de la distribution normale univariée est la **distribution normale multivariée**. La densité jointe de  $n$  variables *indépendantes*  $N(0, 1)$  est simplement le produit de  $n$  densités univariées  $N(0, 1)$ . Ainsi, si  $\mathbf{x}$  est un vecteur de dimension  $n$  d'élément type  $x_i \sim N(0, 1)$ , la densité jointe est

$$f_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x_i^2\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{x}\right). \quad (\text{B.15})$$

Nous notons symboliquement cette densité  $N(\mathbf{0}, \mathbf{I})$ . Le premier argument est un vecteur composé de  $n$  zéros, chacun étant dans ce cas l'espérance de l'élément correspondant de  $\mathbf{x}$ . Le second argument est une matrice identité de dimension  $n \times n$ , qui est dans ce cas la matrice de covariance de  $\mathbf{x}$ . C'est l'exemple le plus simple d'une densité normale multivariée.

Un vecteur aléatoire qui suit n'importe quelle distribution normale multivariée peut se dériver à partir de  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$ . Considérons par exemple un vecteur  $\mathbf{y}$  de  $n$  variables aléatoires issues de combinaisons linéaires des éléments de  $\mathbf{x}$ . Ceci implique  $\mathbf{y} \equiv \mathbf{A}^\top \mathbf{x}$  pour une matrice non singulière non aléatoire  $\mathbf{A}$  de dimension  $n \times n$  quelconque. Il est clair que  $E(\mathbf{y}) = \mathbf{0}$  et que  $\mathbf{V}(\mathbf{y}) = \mathbf{A}^\top \mathbf{A}$ ; voir (B.06). La distribution du vecteur  $\mathbf{y}$  de dimension  $n$  est, par définition, la distribution  $N(\mathbf{0}, \mathbf{A}^\top \mathbf{A})$ . Ainsi nous voyons que, comme pour la distribution  $N(\mathbf{0}, \mathbf{I})$ , l'argument matriciel est la matrice de covariance des

éléments de  $\mathbf{y}$ . Puisque toute matrice de covariance  $\mathbf{V}$  peut s'écrire comme  $\mathbf{A}^\top \mathbf{A}$  pour une matrice  $\mathbf{A}$  appropriée, nous pouvons caractériser la densité  $N(\mathbf{0}, \mathbf{V})$  pour une matrice  $\mathbf{V}$  définie positive quelconque en explicitant la densité jointe de  $\mathbf{y}$ .

La forme la plus générale de la distribution normale multivariée s'obtient à partir du vecteur aléatoire  $\mathbf{y} \sim N(\mathbf{0}, \mathbf{V})$  de dimension  $n$  en lui associant un vecteur  $\boldsymbol{\mu}$  de dimension  $n$ . Puisque  $E(\mathbf{y} + \boldsymbol{\mu}) = \boldsymbol{\mu}$ , l'espérance du vecteur aléatoire ainsi construit est  $\boldsymbol{\mu}$ . Ainsi nous notons symboliquement  $N(\boldsymbol{\mu}, \mathbf{V})$  la distribution normale multivariée générale, avec un vecteur d'espérances  $\boldsymbol{\mu}$  et une matrice de covariance  $\mathbf{V}$ .

Avant de dériver la densité jointe de la distribution  $N(\boldsymbol{\mu}, \mathbf{V})$ , il faut résoudre un problème plus général. Supposons connue la distribution de la variable aléatoire  $x$ , où  $x$  est pour l'instant scalaire. Quelle est alors la distribution d'une autre v.a.  $y$  qui est une fonction déterministe de  $x$ ? Pour faire simple, supposons que  $y = g(x)$  pour une fonction quelconque  $g$  monotone croissante. En terme de la c.d.f., le calcul est immédiat:

$$\Pr(y < Y) = \Pr(g(x) < Y) = \Pr(x < g^{-1}(Y)) = F_x(g^{-1}(Y)).$$

Notons que  $g^{-1}$  existe du fait de l'hypothèse de monotonie de  $g$ . Ainsi la c.d.f. de  $y$  est

$$F_y(Y) = F_x(g^{-1}(Y)). \quad (\text{B.16})$$

Nous pouvons alors déterminer la densité de  $y$  en dérivant (B.16):

$$f_y(Y) = f_x(g^{-1}(Y)) \frac{dg^{-1}(Y)}{dY} = \frac{f_x(g^{-1}(Y))}{g'(g^{-1}(Y))}. \quad (\text{B.17})$$

Ainsi la densité de  $y$  est simplement égale à la densité de  $x$  divisée par la dérivée première de  $g(\cdot)$ , les deux étant évaluées en  $g^{-1}(Y)$ . Les lecteurs peuvent être intéressés par la dérivation de la densité normale univariée générale (B.14) à partir de la densité normale centrée réduite (B.12) en appliquant ce résultat.

Il existe un moyen mnémotechnique simple pour lier les deux formes du résultat (B.17). Il rappelle simplement que

$$f_y(Y) dy = f_x(X) dx.$$

Le moyen mnémotechnique est relayé à une expression mathématique explicite en divisant soit par  $dy$  soit par  $dx$  et en posant  $X = g^{-1}(Y)$  ou  $Y = g(X)$ . La première possibilité conduit à l'expression centrale de (B.17), alors que la seconde conduit à

$$f_y(g(X))g'(X) = f_x(X),$$

qui est l'équivalent de l'expression la plus à droite de (B.17).

Si  $g$  était une fonction monotone décroissante plutôt que croissante, (B.17) resterait vraie si la dérivée  $g'$ , négative, était remplacée par sa valeur absolue  $|g'|$  (le montrer constitue un bon exercice). Si  $g$  n'était pas monotone, il faudrait alors découper son domaine de définition en sous-espaces où elle serait monotone, et (B.17) s'appliquerait à chacun de ces sous-espaces, au moins localement. L'élément clé est qu'une valeur  $Y$  peut à présent correspondre à plusieurs valeurs  $X$ , et dans ce cas la densité de  $y$  en  $Y$  est la somme des contributions calculées en utilisant (B.17) pour chacune des valeurs de  $X$ .

Pour dériver la densité de la distribution normale multivariée, il nous faut trouver une version multivariée de (B.17). Supposons qu'une v.a. vectorielle  $\mathbf{y}$  de dimension  $n$  soit donnée en terme d'une autre v.a. vectorielle  $\mathbf{x}$  de dimension  $n$  par l'application déterministe  $\mathbf{y} = \mathbf{g}(\mathbf{x})$ , que nous supposons bijective. Un argument plus fin que celui utilisé dans le cas scalaire montre que

$$f_{\mathbf{y}}(\mathbf{Y}) = f_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{Y})) |\det \mathbf{J}(\mathbf{Y})|, \quad (\text{B.18})$$

où  $\mathbf{J}(\mathbf{Y})$ , la matrice Jacobienne de la transformation de  $\mathbf{y}$  en  $\mathbf{x}$ , est la matrice de dimension  $n \times n$  des dérivées de  $\mathbf{g}^{-1}(\mathbf{Y})$  par rapport aux éléments de  $\mathbf{Y}$ . La notation  $|\det(\cdot)|$  désigne la valeur absolue du déterminant. La valeur absolue du déterminant apparaît dans (B.18) essentiellement pour la même raison que le cas univarié nécessite la valeur absolue de  $g'$  quand  $g'$  est négative.

Il est souvent commode lors du calcul de déterminant dans (B.18) d'utiliser le fait que la matrice Jacobienne de la transformation de  $\mathbf{y}$  en  $\mathbf{x}$  est l'inverse de la matrice Jacobienne de la transformation de  $\mathbf{x}$  en  $\mathbf{y}$ , ainsi que la propriété que le déterminant de l'inverse d'une matrice est l'inverse du déterminant de la matrice. Ainsi, si  $\mathbf{J}^*$  désigne la matrice Jacobienne  $\mathbf{g}(\mathbf{X})$ , une manière alternative d'écrire (B.18) est

$$f_{\mathbf{y}}(\mathbf{Y}) = f_{\mathbf{x}}(\mathbf{g}^{-1}(\mathbf{Y})) |\det \mathbf{J}^*(\mathbf{Y})|^{-1}.$$

Les lecteurs motivés sont encouragés à travailler en détail la dérivation de (B.18). Celle-ci n'est pas difficile en principe, du moins pour le cas  $2 \times 2$ . Les lecteurs férus de théorie de l'intégration comprendront intuitivement (B.18) en notant que le déterminant est le ratio des volumes infinitésimaux dans les espaces de  $\mathbf{x}$  et de  $\mathbf{y}$ , respectivement; voir l'Annexe A.

Nous pouvons à présent revenir au problème de la détermination de la densité normale multivariée. Supposons que  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$  et  $\mathbf{y} = \mathbf{A}^\top \mathbf{x} + \boldsymbol{\mu}$ . Ceci implique que  $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ , où  $\mathbf{V} \equiv \mathbf{A}^\top \mathbf{A}$ . La matrice Jacobienne de la transformation de  $\mathbf{y}$  en  $\mathbf{x}$  est dans ce cas  $(\mathbf{A}^\top)^{-1}$ . Puisque la densité de  $\mathbf{x}$  est (B.15), le résultat (B.18) implique que la densité de  $\mathbf{y}$  soit

$$\begin{aligned} & (2\pi)^{-n/2} |\det \mathbf{A}|^{-1} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{A}^{-1} (\mathbf{A}^\top)^{-1} (\mathbf{y} - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right), \end{aligned} \quad (\text{B.19})$$

où  $|\mathbf{V}|$  est le déterminant de  $\mathbf{V}$ , toujours positif. La seconde ligne exploite le fait que la matrice de covariance  $\mathbf{V}$  est égale à  $\bar{\mathbf{A}}^T \mathbf{A}$ . (B.19) est le moyen traditionnel d'écrire la densité normale multivariée pour le cas général où  $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ .

De nombreuses distributions bien connues peuvent se définir en terme de la distribution normale centrée réduite. La plus étrange de celles-ci est sans doute la **distribution de Cauchy**. Par définition, c'est la distribution du rapport de deux variables aléatoires normales centrées réduites indépendantes. Soient  $x$  et  $y$  deux telles v.a. La densité jointe de  $x$  et  $y$  est, d'après (B.15),

$$(2\pi)^{-1} \exp\left(-\frac{1}{2}(x^2 + y^2)\right).$$

Pour dériver la densité de Cauchy, nous devons effectuer un changement de variables vers des coordonnées polaires  $r$  et  $\theta$  qui correspondent à  $x$  et  $y$ . La relation entre ces coordonnées polaires et les coordonnées cartésiennes est

$$\begin{aligned} x &= r \cos \theta; & y &= r \sin \theta; \\ r &= (x^2 + y^2)^{1/2}; & \theta &= \tan^{-1}(y/x). \end{aligned}$$

Le déterminant de la matrice Jacobienne de la transformation de  $(r, \theta)$  en  $(x, y)$  est  $r \sin^2 \theta + r \cos^2 \theta = r$ . Par conséquent, la densité jointe de  $r$  et  $\theta$  est

$$(2\pi)^{-1} r e^{-r^2/2}. \quad (\text{B.20})$$

Celle-ci ne dépend aucunement de  $\theta$ , ce qui implique que la densité de  $\theta$  doit être uniforme sur un intervalle quelconque. A l'évidence, puisque  $\theta$  est un angle exprimé en radians, cet intervalle doit être  $[0, 2\pi]$ . Nous pouvons montrer ce résultat plus formellement en intégrant (B.20) par rapport à  $r$  sur l'intervalle allant de 0 à  $\infty$ . Le résultat, qui est la densité de  $\theta$ , est simplement  $(2\pi)^{-1}$ . Ceci est en fait la densité d'une variable aléatoire uniformément distribuée sur l'intervalle  $[0, 2\pi]$ .

La variable aléatoire de Cauchy  $z \equiv y/x$  est reliée à  $\theta$  par la relation  $z = \tan \theta$ . La matrice Jacobienne (scalaire ici), de la transformation de  $z$  en  $\theta$  est par conséquent l'inverse de la dérivée de  $\tan \theta$  par rapport à  $\theta$ . Cette dérivée est  $\sec^2 \theta$ . Avant d'expliciter la densité de  $z$ , il faut remarquer que, lorsque  $\theta$  varie de 0 à  $2\pi$ , chaque valeur de  $z$  est générée exactement deux fois, puisque  $\tan(\pi + \theta) = \tan \theta$ . Ainsi nous concluons que la densité de la distribution de Cauchy est

$$2(2\pi)^{-1} \frac{1}{\sec^2 \theta} = \frac{1}{\pi(1 + \tan^2 \theta)} = \frac{1}{\pi(1 + z^2)}.$$

Il est clair que si nous essayons d'évaluer l'espérance d'une v.a. de Cauchy, nous serons confrontés à l'intégrale

$$\int_{-\infty}^{\infty} \frac{z dz}{\pi(1 + z^2)},$$



qui diverge pour les deux bornes d'intégration. Ainsi la distribution de Cauchy ne possède aucun moment.

La **distribution chi-deux** est d'une importance encore plus grande pour les économètres que la distribution de Cauchy. La distribution dépend de deux paramètres, un entier positif, appelé nombre de **degrés de liberté**, et un réel positif, appelé **paramètre de non centralité**, ou **NCP**. L'écriture symbolique d'une variable aléatoire du chi-deux à  $n$  degrés de liberté et de NCP  $\Lambda$  est  $\chi^2(n, \Lambda)$ . Lorsque le NCP est nul, comme c'est souvent le cas, la variable suit la **distribution du chi-deux centrée**. Celle-ci est souvent notée  $\chi^2(n)$  plutôt que  $\chi^2(n, 0)$ .

La distribution du chi-deux centrée est définie au moyen d'un vecteur  $\mathbf{x}$  de dimension  $n$  distribué suivant la  $N(\mathbf{0}, \mathbf{I})$ . Alors la variable aléatoire  $y$  définie comme  $\mathbf{x}^\top \mathbf{x}$  possède une distribution du  $\chi^2(n)$ . Il est clair que  $y$  est la somme au carré de  $n$  v.a. normales centrées réduites indépendantes. Il n'est pas difficile de calculer la densité de  $\chi^2(n)$  à l'aide de cette remarque, à condition de maîtriser les coordonnées polaires en dimension  $n$ . Heureusement, nous n'utilisons pas explicitement cette densité, de sorte que nous éviterons la manipulation. Il est utile de noter que  $E(y) = n$  et  $V(y) = 2n$ .

Lorsque le NCP est non nul, la v.a. suit la **distribution du chi-deux non centrée**. Une variable aléatoire suivant la distribution du  $\chi^2(n, \Lambda)$  peut se construire comme la somme des carrés de  $n - 1$  v.a. normales centrées réduites indépendantes, plus le carré d'une autre v.a. indépendante des autres, distribuée suivant la  $N(\Lambda^{1/2}, 1)$ . Il peut aussi se construire comme la somme de  $n$  v.a. indépendantes  $x_i$  au carré, où  $x_i \sim N(\mu_i, 1)$  et  $\Lambda = \sum_{i=1}^n \mu_i^2$ . La première définition est à l'évidence un cas particulier de la seconde. La démonstration que la densité ne dépend que de la somme  $\sum_{i=1}^n \mu_i^2$  et non pas des  $\mu_i$  individuels dépasse les objectifs de cette annexe.

La distribution du chi-deux non centrée possède la propriété suivante. Pour tout nombre positif  $c$ ,

$$\Pr(\chi^2(n, \Lambda) > c)$$

est une fonction croissante de  $n$  et de  $\Lambda$ . Ce résultat se démontre aisément. Ce n'est pas le cas d'un résultat de Das Gupta et Perlman (1974) (la démonstration utilise des techniques qui dépassent de loin le niveau de cet ouvrage). Ce résultat est au coeur des arguments traitant de la puissance des tests basés sur des statistiques ayant asymptotiquement la forme du chi-deux. Il est comme suit. Pour tout  $\alpha \in [0, 1]$ , supposons que  $c_{n\alpha}$  satisfasse la condition  $\Pr(\chi^2(n) > c_{n\alpha}) = \alpha$ . Ainsi  $c_{n\alpha}$  est la valeur critique pour un test de niveau  $\alpha$  utilisant la distribution du chi-deux centré à  $n$  degrés de liberté. Alors, pour chaque NCP  $\Lambda$ ,

$$\Pr(\chi^2(n, \Lambda) > c_{n\alpha})$$

est une fonction croissante de  $n$ . Ainsi, pour un NCP donné, la puissance de test diminuera lorsque le nombre de degrés de liberté augmentera.

De nombreuses statistiques de test sont calculées comme une forme quadratique composée d'un vecteur de v.a. (asymptotiquement) distribuées normalement et d'une estimation de l'inverse de leur matrice de covariance. Ces statistiques de test sont asymptotiquement distribuées suivant un chi-deux centré. Ce résultat dépend du fait que si un vecteur  $\mathbf{x}$  de dimension  $n$  est distribué suivant la  $N(\mathbf{0}, \mathbf{V})$ , la forme quadratique  $z \equiv \mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x}$  obéit à la distribution  $\chi^2(n, 0)$ . De fait, par souci d'économie, nous démontrons le résultat plus général que si  $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$ ,  $z$  sera distribuée suivant la  $\chi^2(n, \boldsymbol{\mu}^\top \mathbf{V}^{-1} \boldsymbol{\mu})$ .

Soit  $\boldsymbol{\eta}$  une matrice symétrique telle que  $\mathbf{V}^{-1} = \boldsymbol{\eta} \boldsymbol{\eta}^\top$ , et considérons le vecteur aléatoire  $\mathbf{y} \equiv \boldsymbol{\eta} \mathbf{x}$ . Nous avons construit  $\mathbf{y}$  de sorte que  $\mathbf{y}^\top \mathbf{y} = \mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x} = z$ . Le vecteur  $\mathbf{y}$  est manifestement normal multivarié, d'espérance  $\boldsymbol{\eta} \boldsymbol{\mu}$  et de matrice de covariance  $\boldsymbol{\eta} \mathbf{V} \boldsymbol{\eta}^\top = \mathbf{I}$ . Par la seconde définition de la distribution du chi-deux non centrée,  $z$  doit être distribuée suivant la  $\chi^2(n, \boldsymbol{\mu}^\top \mathbf{V}^{-1} \boldsymbol{\mu})$ , comme requis. Le résultat selon lequel  $z \sim \chi^2(n)$  pour le cas particulier  $\boldsymbol{\mu} = \mathbf{0}$  découle immédiatement de ce résultat plus général.

Un résultat étroitement relié est le suivant. Supposons que  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_n)$ . Alors, si  $\mathbf{P}$  est une matrice de projection orthogonale de dimension  $n \times n$  de rang  $r < n$ , la **forme quadratique idempotente**  $\mathbf{x}^\top \mathbf{P} \mathbf{x}$  est distribuée suivant la  $\chi^2(r)$ . Pour le comprendre, il est pratique d'exprimer la matrice  $\mathbf{P}$  sous la forme  $\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$ , pour une matrice  $\mathbf{Z}$  adéquate quelconque de dimension  $r \times n$  telle que  $\mathcal{S}(\mathbf{Z}) = \mathcal{S}(\mathbf{P})$ . Alors

$$\mathbf{x}^\top \mathbf{P} \mathbf{x} = \mathbf{x}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{x}.$$

Evidemment, le vecteur  $\mathbf{Z}^\top \mathbf{x}$  de dimension  $r$  suit la distribution  $N(\mathbf{0}, \mathbf{Z}^\top \mathbf{Z})$ . Par conséquent,  $\mathbf{x}^\top \mathbf{P} \mathbf{x}$  est une forme quadratique composée d'un vecteur normal multivarié de dimension  $r$  et de l'inverse de sa matrice de covariance. Le résultat recherché provient immédiatement des résultats des paragraphes précédents.

La **distribution  $F$**  peut se définir en terme de deux variables aléatoires indépendantes, chacune obéissant à une distribution du  $\chi^2$ . Puisqu'aucune, une seule, ou les deux v.a. peut(peuvent) être non centrée(s), la distribution  $F$  peut être centrée, non centrée, ou doublement non centrée. La **distribution  $F$  centrée** à  $n$  et  $d$  degrés de liberté (pour "numérateur" et "dénominateur") est la distribution du rapport de deux v.a. du  $\chi^2$  centrés indépendantes à  $n$  et  $d$  degrés de liberté respectivement, chacune étant divisée par le nombre de ses degrés de liberté. Symboliquement,

$$F(n, d) = \frac{\chi^2(n)/n}{\chi^2(d)/d}.$$

La **distribution  $F$  non centrée** à  $n$  et  $d$  degrés de liberté et un NCP  $\Lambda$  est la distribution du rapport d'un numérateur distribué suivant  $n^{-1} \chi^2(n, \Lambda)$  et d'un dénominateur qui lui est indépendant distribué suivant  $d^{-1} \chi^2(d, 0)$ . La

**distribution  $F$  doublement non centrée** à  $n$  et  $d$  degrés de liberté et des NCP  $\Lambda_n$  et  $\Lambda_d$  est la distribution du rapport d'un numérateur distribué suivant  $n^{-1}\chi^2(n, \Lambda_n)$  et d'un dénominateur qui lui est indépendant distribué suivant  $d^{-1}\chi^2(d, \Lambda_d)$ . Les densités de ces deux distributions  $F$  sont connues et tabulées—consulter, par exemple, Abramowitz et Stegun (1965)—mais ne sont pas d'un grand intérêt pour les économètres. Dans la pratique, nous n'avons besoin que d'un programme de calcul de la c.d.f. et de l'inverse de la c.d.f. de la distribution  $F$  centrée, et de tels programmes sont disponibles dans la plupart des bons progiciels de statistique.

Enfin, nous abordons la **distribution de Student**, qui est souvent simplement dénommée **distribution  $t$** . La distribution de Student à  $n$  degrés de liberté est notée  $t(n)$  et définie comme la distribution d'une v.a. normale centrée réduite divisée par une v.a. qui lui est indépendante distribuée selon la racine carrée de  $n^{-1}\chi^2(n, 0)$ . Evidemment, le carré d'une variable aléatoire distribuée suivant une  $t(n)$  est distribué suivant une  $F(1, n)$  centrée. Etant donnée la définition de la distribution du chi-deux centrée, il est clair que la loi des grands nombres peut s'appliquer à  $n^{-1}\chi^2(n, 0)$  quand  $n \rightarrow \infty$ . Puisque l'espérance de chaque variable normale centrée réduite au carré dans la définition est égale à un, la limite de  $n^{-1}\chi^2(n, 0)$  doit être 1. Par conséquent, la distribution  $t(n)$  tend vers la distribution normale centrée réduite lorsque  $n \rightarrow \infty$ .

Pour la plupart des valeurs de  $n$ , la distribution  $t$  ressemble énormément à la distribution normale centrée réduite, mais possède des queues de distribution légèrement plus épaisses. La différence entre la distribution  $t$  et la distribution normale centrée réduite est très faible pour  $n \geq 100$ ; par exemple, la valeur critique à 5% d'un test bilatéral est 1.960 pour  $N(0, 1)$  et 1.984 pour  $t(100)$ . Cependant cette différence peut s'accroître pour des valeurs très faibles de  $n$ . La distribution de  $t(1)$  est évidemment la même que la distribution de Cauchy, et elle ne possède par conséquent aucun moment. La distribution  $t(2)$  possède un premier moment nul mais n'a pas de moment d'ordre supérieur. En général, la distribution  $t(n)$  possède des moments jusqu'à l'ordre  $n - 1$ .

A l'occasion, la **distribution  $t$  non centrée** survient. Elle est définie comme

$$t(n, \mu) = \frac{N(\mu, 1)}{(n^{-1}\chi^2(n, 0))^{1/2}}.$$

Le NCP est  $\mu$ , et le carré d'une telle variable aléatoire est distribué suivant une  $F$  non centrée à 1 et  $n$  degrés de liberté et un NCP  $\mu^2$ .

Pour davantage de détails sur les propriétés des distributions discutées dans cette section, les lecteurs peuvent consulter Kendall et Stuart (1977) ou Johnson et Kotz (1970a, 1970b).

## TERMES ET CONCEPTS

covariance	fonction de distribution jointe
degrés de liberté	fonction intégrante
distribution de Cauchy	fonction de logvraisemblance
distribution $F$ , centrée, non centrée, et doublement non centrée	forme quadratique idempotente
espace des événements, ou espace des réalisations	inégalité de Chebyshev
espérance	inégalité de Jensen
densité jointe	indépendance statistique
densité marginale	intégrale de Stieltjes
distribution de probabilité	matrice de covariance
distribution centrée réduite	mesure de probabilité
distribution de Student, centrée et non centrée	mesure de tendance centrale
distribution du chi-deux, centrée et non centrée	moments d'ordre un, deux, trois, et supérieur
distribution marginale	moments des variables aléatoires, centrés et non centrés
distribution normale univariée	moyenne, de population et d'échantillon
distribution normale multivariée	normalisation (d'une densité)
écart standard	paramètre de non centralité, ou NCP
écart type	queues de distribution, de droite et de gauche
espace de probabilité	sigma-algèbre
événement composite	support d'une densité
fonction affine	variable aléatoire
fonction de densité de probabilité, ou p.d.f.	variable aléatoire scalaire
fonction de distribution, ou c.d.f.	variable aléatoire vectorielle
fonction de répartition	variance